

DOI <https://doi.org/10.15407/usim.2018.02.051>

УДК 004.65:004.7:004.75:004.738.5

А.А. УРСАТЬЕВ, канд. техн. наук, Международный научно-учебный центр информационных технологий и систем НАН Украины и МОН Украины, просп. Глушкова, 40, Киев 03187, Украина, aleksei@irtc.org.ua

БОЛЬШИЕ ДАННЫЕ. АНАЛИТИЧЕСКИЕ БАЗЫ ДАННЫХ И ХРАНИЛИЩА: *TERADATA*

Статья представляет собой продолжение исследований Больших Данных и инструментария, трансформируемого в новое поколение технологий и архитектур платформ баз данных и хранилищ для интеллектуального вывода. Рассмотрен ряд прогрессивных разработок известных в мире ИТ-компаний, в частности БД Teradata.

Ключевые слова: архитектура MPP, логические хранилища данных LDW, платформа Warehouse Appliance, SN (Shared Nothing)-архитектура, Teradata Aster Discovery — платформа обнаружения, Teradata IntelliCloud™ — данные и аналитика по модели SaaS, Teradata QueryGrid™ — высокопроизводительный связующий элемент между системами в гетерогенных аналитических средах, унифицированная архитектура данных Teradata™ (UDA).

***Teradata* (NYSE: TDC). Общая характеристика**

Прообразом системы, ныне широко известной, стала параллельная аналитическая база данных, созданная специально для поддержки принятия решений в банковской сфере. Такая архитектура *Massively Parallel Processing* (MPP) впервые была представлена компанией *Teradata* в конце 70-х годов XX ст. на нескольких микропроцессорах. Название *Teradata* символизировало желаемую возможность управления терабайтами данных. Сегодня предложения *Teradata* — ведущей компании аналитических решений — содержат СУБД, масштабируемые высокопроизводительные экосистемы для бизнес-аналитики, охватывающие гибридные облачные решения, хранилища данных и решения облачных хранилищ как на собственном управляемом облаке, так и на инфраструктуре общественного облачного провайдера. Сфера ее интересов — финансовые услуги, производство, связь, розничная

торговля, здравоохранение, СМИ, индустрия развлечений и др. [52].

Как следует из [53], несмотря на неутешительные прогнозы о неопределенном будущем параллельных машин баз данных, параллельные системы БД начинают вытеснять традиционные компьютеры основного класса, так как они позволяют работать со значительно более крупными БД в режиме, поддерживающем транзакции. Концепция *Teradata* — высокопараллельные, основанные на языке *SQL* системы без совместного использования ресурсов на базе обычных, массово производимых процессоров, дисков и памяти. При такой архитектуре (*SN*, *Shared Nothing*¹) каждая память и диск(и) находятся в распоряжении какого-либо процессора, работающего как сервер хранящихся в них данных. Реализован ди-

¹ *Shared Nothing* — распределенная вычислительная архитектура, в которой каждый узел — независимый и самодостаточный, для возникновения конфликта нет ни одной точки соприкосновения.

зайн, отличный от главенствующей на тот момент архитектуры мэйнфреймов, — объединение в высокоскоростную сеть вычислительных модулей, работающих параллельно и независимо друг от друга. По сети архитектуры *SN* передаются только запросы и ответы. Непосредственные обращения к памяти и дискам обрабатываются локальным процессором, и только отфильтрованные данные передаются запрашивающей программе. Это позволяет реализовать более расширяемую архитектуру путем минимизации трафика в соединительной сети. *Teradata* использует избыточную древовидную соединительную *Y*-сеть. Такое решение демонстрирует почти *линейные ускорение* и *расширяемость*² при сложных нагрузках и при транзакционной обработке запросов, а также значительное превосходство по скорости над традиционными серверами основного класса при обработке больших (терабайтных) баз данных. Загрузка, архивирование и обработка данных выполняются также параллельно [52, 53].

Существующие реляционные подходы *Teradata* обеспечивают полное соблюдение *ACID* (*Atomicity, Consistency, Isolation, Durability* — атомарность, согласованность, изолированность, надежность). Помимо устойчивости в разделении данных по сети вследствие *SN*-архитектуры, в *Teradata* существует ряд журналов, таких как *Recovery³ Journals* (*Transient Journal, Down-AMP Recovery Journal*), которые, не жертвуя полноценными *ACID* транзакциями в пользу *eventual consistency* (согласованность в итоге), препятствуют нарушению требований *ACID*. Так, атомарность поддерживается транзакционными механизмами и журналом *Transient Journal*, который обеспечивает возмож-

ность отката транзакций. В последнем случае состояние восстанавливается [54].

Высказывание «*SQL* не масштабируется»⁴ никогда не было истинным или релевантным. Не существует технологии любого типа, которая масштабируется при ее использовании некорректно. Понимаемая трудность масштабирования реляционных БД (РБД) обусловлена ее структурированным хранением, низкой избыточностью данных и, главное, соблюдением *ACID* и связанными механизмами блокировки. Плата за использование РБД в качестве платформы с практически неограниченной масштабируемостью — их сложность. Горизонтальное масштабирование в РСУБД затруднено. Среди существующих методов принято выделять репликацию, партиционирование⁵ (секционирование) и шардинг⁶ (*Sharding*) — одна из форм разделения. Большинство СУБД имеют шардинг, используя несколько машин для горизонтального масштабирования. Технология *Teradata* с 80-х годов XX ст. реализовала методы «физического очертания» (*sharding methods to support large-scale DBs*) для поддержки ширококомасштабных БД (размер «терабайта») в то время, когда несколько сотен мегабайт были большими коммерческими данными.

Все же *Teradata* специально разработана для хранилища данных, и наибольший эффект параллелизма архитектуры *SN* проявляется при агрегированных запросах, характерных для аналитической обработки. Она отлично подходит для сбора множества данных и

² Идеальная параллельная система обладает двумя главными свойствами: *линейное ускорение* — вдвое большее аппаратное обеспечение выполнит ту же задачу в два раза быстрее; *линейная расширяемость* — вдвое большее аппаратное обеспечение выполнит вдвое большую задачу за то же время.

³ *Teradata Wiki* — <http://www.teradatawiki.net/2013/09/Teradata-Journals.html>

⁴ *Karwin Bill*. Why are relational/SQL systems difficult to scale? — Mar 22, 2017 — <https://www.quora.com/Why-are-relational-SQL-systems-difficult-to-scale>

⁵ Партиционирование (секционирование) — это разделение набора данных на отдельные части.

⁶ Шардинг — разделение данных на уровне ресурсов. Полученные в результате разделения объекты разносятся по разным серверам БД. В реляционных БД инструментарий, используемый администратором БД для осуществления шардирования, доступен только в специальных версиях. Такое ограничение в РБД вполне логично, предоставляется выбор: что важнее — поддержка транзакций в неразделенной среде или поддержка устойчивости к разделению по сети с согласованностью в конечном счете.

предназначена для приложений хранения, способных обрабатывать большие объемы данных, и обслуживания нескольких одновременных пользователей. Система перемещает данные в хранилище данных, из которого их можно вызвать и проанализировать. В таком случае это хороший инструмент. Для решений, не связанных с агрегацией данных, система менее эффективна.

Teradata — высоко параллельная, производительная система, масштабируемая до петабайт данных, с распределением наборов записей (кортежей)⁷ на основе хэширования на подмножестве узлов (*AMP*) с использованием собственного хэш-алгоритма обработки полей (атрибутов). Кортёжи размещаются в зависимости от значения хэш-функции, примененной к атрибуту каждого кортежа. Функция определяет конкретный диск, на котором будет размещен кортеж [53, 54].

Разделение с хэшированием идеально подходит для прикладных программ, которым требуется только последовательный и ассоциативный доступ к данным. Хэширование используется для расщепления вывода реляционных операторов в промежуточные отношения. Каждый оператор полностью выполняется во всех участвующих узлах, прежде чем начинается выполнение следующего оператора — альтернативы конвейерного параллельного выполнения при обработке запроса [53].

Логическая архитектура РСУБД *Teradata* представлена схемой узла (рис. 5), где: *AMP* — *Access Module Processor*, обеспечивающий извлечение данных с ассоциированного с ним диска (дисков); *PE* — *Parsing Engine*, ответственный за контроль сессии и обработку запросов пользователя; интерфейсные процессоры поддерживают связь с главным сервером, осуществляют синтаксический разбор и оптимизацию запросов, а также координацию

AMP во время выполнения запросов. *BYNET* — среда обмена сообщениями между компонентами системы. *PE* и *AMP* объединяются термином виртуальный процессор *VPROC* [54].

Для балансировки нагрузки между модулями *AMP* и другими административными задачами используются средства *Teradata Manager*, *DBSConsole* и *Teradata Administrator*. В частности они позволяют задавать «фильтры» и приоритеты для выполняемых на *AMP*, либо сервере в целом, пользовательских процессов.

Все данные обрабатываются при выполнении пользовательских запросов параллельно и независимо друг от друга. Для максимальной эффективности обработки запросов данные должны быть распределены между *AMP* равномерно. *BYNET* отвечает за коммуникацию между компонентами системы и за высокоскоростную передачу данных между *AMP*. В платформах *Warehouse Appliance* этот высокопроизводительный компонент реализован поверх *Ethernet*, а в *Active Enterprise Data Warehouse* — это отдельный аппаратный модуль, так как при таких объемах *Ethernet* уже не эффективен.

В СУБД *Teradata* нет хинтов⁸ оптимизатора, его работа основана на статистической информации о данных. Зная это обстоятельство и по каким критериям пользователь их запрашивает, оптимизатор выбирает способы извлечения данных. Оптимизатор запросов принимает решение, основываясь на доступной объективной информации (количество узлов и *AMP* в системе, количество и типы процессоров, памяти в данный момент, типы дисков и демографию данных). Демографическая информация — статистика — включает в себя количество строк в таблице и средний размер строки, строк с одним и тем же значением колонки и пр. Знание этих параметров также позволяет оптимизатору правильно рассчитывать размеры временной памяти (*spool*), выделяемой запросу для проведения преобра-

⁷ Реляционная база данных состоит из отношений (*relation*) или файлов, которые в свою очередь содержат кортежи (*tuples*) или записи. Все кортежи в отношении содержат один и тот же набор атрибутов (поля в терминологии языка *COBOL*) [53].

⁸ Подсказка в языке *SQL* (англ. *hint*) — средство, позволяющее явным образом влиять на план запроса.

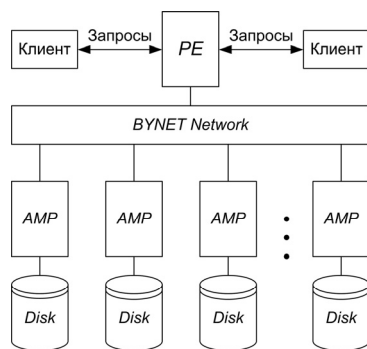


Рис. 5

зований данных [55]. Оптимизатор⁹ обеспечивают надежные результаты независимо от сложности *SQL*-запросов.

Общие базы данных *MPP* не поддерживают индексы. Лучше всего они подходят для хранения данных и аналитических рабочих нагрузок, а традиционные индексы *B-Tree*¹⁰ редко бывают полезны в этих ситуациях. Архитектура *MPP*, равномерное распределение данных по узлам системы делают их обработку эффективной и в отсутствие индексов — дополнительных элементов обеспечения производительности. Эта возможность позволяет *Teradata* легко справляться с *ad-hoc* запросами, не предусмотренными разработчиками физической модели, но для смешанной нагрузки, когда в системе одновременно выполняются и тактические (*OLTP*-подобные), и стратегические (*OLAP*, *DSS*, *Data Mining*) [18] запросы, этого явно недостаточно. Создание индексов для повышения производительности тактических запросов, которые имеют достаточно высокие требования к времени отклика и пропускной способности, критерии их выбора и техник физического моделирования, механизмы работы СУБД *Teradata* с данными с учетом ее параллельной архитектуры приведены в [54, 56, 57].

Вместе с тем, в комментариях [54] высказывается мнение, что вряд ли СУБД *Teradata* может быть одинаково хорошо приемлема для абсолютно всех задач. При оперативной обработке транзакций, СУБД с *MPP*-архитектурой будет сложно конкурировать с транзакционными системами, несмотря на все предпринимаемые разработчиками меры повышения производительности. *Teradata* — это в первую очередь быстрая обработка и анализ больших объемов, от сотен терабайт до петабайт, данных. Кастомная аналитика, отчеты, аналитические *CRM*, *Data Mining* — лишь небольшой список задач, хорошо решаемых с помощью СУБД *Teradata* [52, 54].

Teradata предоставляет оптимизированное использование твердотельных накопителей (*SSD*) и жестких дисков (*HDD*) в среде хранилищ объемом на петабайт данных. Оптимизацию обеспечивает *Teradata Virtual Storage*TM (*TVS*) — интеллектуальное программное обеспечение управления гибридной средой хранения, автоматически перемещающее наиболее часто используемые данные на быстрые *SSD*. В основе эффективности гибридных систем хранения то, что разные данные используются потребителями с неодинаковой частотой: данные, лежащие мертвым грузом, считают холодными, а наиболее востребованные — горячими. Управление по «температуре данных» происходит по приоритету выделения системных ресурсов на основе бизнес-правил с использованием возможностей устройств хранения [58].

Версия *Teradata Database* 13.0 лидирует на рынке с автоматической виртуальной памятью, увеличенной на 30 процентов производительностью и рядом усовершенствований, поддерживающих активное хранилище данных [18, 52].

Teradata Database версии 14 (2012 г.) хранит данные как в виде строк, так и в виде столбцов (горизонтальное и вертикальное партиционирование¹¹). Возможно гибридное хра-

⁹ How is Teradata different from Oracle, given that both are RDBMS databases? — <https://www.quora.com/How-is-Teradata-different-from-Oracle-given-that-both-are-RDBMS-databases>

¹⁰ Обзор типов индексов *Oracle*, *MySQL*, *PostgreSQL*, *MS SQL* — 27 авг 2010 — <https://habrahabr.ru/post/102785/>

¹¹ Горизонтальное партиционирование — разделение данных по строкам, когда разные части записей хранятся в отдельных секциях. Вертикальное партиционирование — разделение данных по столбцам.

нение данных в *DB Teradata: Teradata Columnar*, при котором таблицы одновременно могут использовать два метода партиционирования. Особенность такого подхода в том, что партиции по строкам и партиции по колонкам можно использовать одновременно в одной и той же таблице, т.е. — партиции внутри партиций. Выбираются нужные партиции по колонкам, затем только внутри них читаются нужные партиции по строкам.

В помощь физическому проектированию БД, имеется специальный инструмент — *Columnar Analysis Tool*. Он анализирует использование той или иной таблицы *SQL*-запросами и выдает рекомендации по применимости формата хранения по колонкам. Применяется эффективное автоматическое сжатие с использованием механизма автоматической компрессии [52].

Новые технологии обработки

Рост Больших Данных, вызванных, в частности, новыми медиа-источниками, такими как социальные медиа, побудили *Teradata* к приобретению в 2011 г. *Aster Data Systems*, лидирующую в области управления и анализа различными неструктурированными и частично структурированными данными. Большие данные привели к тому, что многие традиционные компании хранилищ данных обновили свои продукты и технологии.

Teradata Aster Discovery — платформа обновления. Аналитические возможности этих двух систем с *Aster's* запатентованными технологиями — *Aster SNAP Framework™* и *SQL-MapReduce®* — по сути, были объединены в двухуровневой платформе *Teradata Aster Discovery* [59]. Эта платформа, содержащая БД *Teradata Aster* и портфолио *Discovery*, значительно повышает способность выполнять процесс выявления «нужных» данных (сбор данных — их подготовка — анализ — визуализация), обуславливающих важнейшие неинтуитивные бизнес-идеи из всех доступных типов данных.

Портфолио *Discovery* [59—61] представляет собой набор готовых к использованию анали-

тических функций *SQL*, *SQL-MapReduce®* и *Graph* для исследования Больших Данных. Содержит модули сбора и подготовки данных, модуль средств аналитики и визуализации (рис. 6). Модуль сбора данных обеспечивает доступ к мультиструктурированным данным, хранящимся в *Apache™ Hadoop™*, *Teradata Data Warehouse* и других реляционных СУБД (*RDBMS*). Модуль подготовки данных содержит адаптеры и преобразователи, позволяющие анализировать и интерпретировать содержимое блогов, *XML*-документов, электронных писем и журналов устройств. Примерный набор функций предоставляет такие возможности: фильтры для удаления выбросов значений данных; *Apache Log Parser* для поддержки пользовательских форматов журналов с определенным пользователем веб-серверов *Apache*; *XML* и *JSON 12 Parsers* для анализа и подготовки *XML*-журналов, создаваемых такими приложениями как веб-журналы и кассовые журналы в розничных магазинах и торговых точках, и др.

Функции преобразования данных предусматривают сессионизацию (*sessionization*) и распаковку форматов для преобразования сложных, неструктурированных данных в значимые форматы для аналитики [61].

Модуль средств аналитики в дополнение к упомянутым пользовательским аналитическим функциям обеспечивает широкий набор готовых к использованию функций для временных рядов, статистических методов, аналитики текста, *Graph* и *SQL*-анализа, таких как *Page-Rank*, *nPath™*, *Confusion Matrix* (матрица ошибок), *Attensity™ ASAS* и др. [59].

Аналитическая функция *SQL-MapReduce* реализована путем слияния двух технологий обработки: открытого и производительного *SQL* и хорошо масштабируемого на множество узлов кластера *Hadoop MapReduce*. Напомним, *Hadoop* — де-факто стандарт технологий для работы с Большими Данными, *MapReduce* —

¹² *Teradata Database* может обрабатывать и хранить записи *JSON* как документ *JSON* или хранить *JSON*-записи в реляционной форме.

Предустановленные приложения		Пакетные решения		Инструменты бизнес-аналитики		Решения бизнес-аналитики
AppCenter						
Сбор данных	Подготовка данных	Средства аналитики	Визуализация	Настраиваемые модули		Встроенные многожандровые масштабируемые аналитические функции
Teradata Hadoop RDBMS Splunk	Адаптеры Преобразователи	Graph Time Series Path and Pattern Text Machine Learning Statistics	• Потока • Иерархии • Аффинити	 		
SQL	SQL-MapReduce®	SQL-GR™	R	...		Обработки аналитики
Aster Database Raw, Column and Aster File Store		Hadoop HDFS Hortonworks and Cloudera		...		Хранение данных

Рис. 6

модель программирования, объединяющая формирование наборов из больших данных на узлах кластера с их обработкой, служит для организации параллельных распределенных вычислений [40, 62]. Преимущества такого конвергентного подхода неоспоримы, если нужна масштабируемость по требованию. Обработка реляционных данных на принципах *MapReduce* обладает большей гибкостью в сравнении с реляционной СУБД с массовым параллелизмом на основе *SQL*-запросов.

SQL-MapReduce® — это инфраструктура (*frameworks*), позволяющая разработчикам писать мощные и выразительные функции *SQL-MapReduce* в языках *Java*, *C#*, *Python*, *C++*, *R* и встраивать их в платформу *Discovery* для высокопроизводительной аналитики. Функции *MapReduce* плавно интегрируются в *SQL*-запросы. Аналитики могут вызывать функции *SQL-MapReduce* с использованием стандартного *SQL* или *R* посредством базы данных *AsterDB* — первый уровень платформы аналитического вывода. Это позволяет встроенным приложениям полностью интегрироваться в механизм *DB Aster Teradata*, которая зарекомендовала себя как платформа для обеспечения сверхбыстрого глубокого анализа массивов данных [59—61].

SQL-GR™ — эта программная конструкция обработчика обеспечивает масштабируемость, итеративную обработку данных и содержит предварительно разработанные функции графа, вызываемые через *SQL*.

Интерфейс *R* для *Aster Analytics*. *R* — мультипарадигмальный язык¹³ с открытым исходным кодом для статистического и интеллектуального анализа данных и работы с графикой [46], содержит встроенные статистические и графические инструменты, включая классификацию, кластеризацию и пр. Возможности базового языка легко расширяются множеством пакетов, благодаря которым он обеспечивает многомерную расширенную аналитику. Однако будучи мощным, гибким и расширяемым, язык *R* несет в себе различные ограничения памяти, данных и обработки. Поэтому *Teradata* предоставляет интерфейс *R* для *Aster Analytics* — решение, которое обеспечивает масштабируемое *R* для бизнес-аналитики. Интеграция программного решения ядра вычислителя (*engine*) с открытым исходным кодом *R* гарантирует единую мощную аналитическую среду для сообщества аналитиков, включая программистов *R*, бизнес-аналитика и научного сотрудника. Пакеты *Aster R* предоставляют интерфейс для функций *Aster Analytics*, которые можно комбинировать с пакетами *R* [61].

База данных *Teradata Aster* с архитектурой *MPP* (см. рис. 6) поддерживает полиморфную модель хранения реляционных данных (*row/column data store*) с полностью неструктурированным хранилищем, размещенным на файловой системе *AFS* (*aster file system*). Данные могут перемещаться бесшовно из реляционного в неструктурированное хранилище на *AFS* и обратно. *DB Aster* включает в себя три аналитических механизма (*SQL*, *MapReduce* и *Graph*), предназначенные для обеспечения оптимальной обработки аналитических задач в больших объемах данных. Аналитика выполняется непосредственно в БД (*in database analytics*), чтобы исключить перемещение данных и использовать распределенную параллельную обработку. *DB Aster* может быть легко расширена для поддержки других аналитиче-

¹³ Основа системы *R* — интерпретируемый язык, поддерживающий структурный, модульный и объектно-ориентированный стили программирования.

ских механизмов на основе языков *R* или *Python*, например *Snap Framework*[™] для расширенной аналитики. Помимо этого, аналитики, предпочитающие *R*, могут создавать *R*-скрипты для реализации многокритериальной расширенной аналитики, сочетая гибкость языка *R* с *Aster's* расширенными возможностями аналитики [61].

В *Teradata Aster* используется *SQL-MapReduce*[®] [59] для выяснения вопросов, ранее невозможного. Это готовые к использованию аналитические функции: цифровая оптимизация маркетинга, анализ социальной сети, обнаружение мошенничества и др. Так, например, крупная компания онлайн-игр сократила время цикла «аналитики мошенничества» с одной недели до 15 мин. путем снижения времени ответа на запрос в 60 раз. Онлайн и социальные медиа-компании, используя функцию *SQL-MapReduce nPath*, анализируют путь пользователя по истории его запросов для улучшения результатов поиска; в целях понимания взаимосвязей и усовершенствования сетей сообществ и их пользователей, применяют ту же функцию для анализа графа (*Graf Analysis*); социальная сеть разрывает пользовательскую сессию при превышении времени активности в сети, улучшая тем самым производительность, и др.

Teradata Aster Snap Framework[™] [61] для расширенной аналитики (рис. 7) — *Aster Analytics*. Полная обработка сетевой аналитики (*Seamless Network Analytics Processing, SNAP*) позволяет, используя унифицированный *SQL*-интерфейс, вызывать одним *SQL*-запросом расширенную аналитику нескольких видов (граф, путь / шаблон, текст, *SQL* и статистический прогнозный анализ). Интегрированный оптимизатор обрабатывает этот «мультижанровый»¹⁴ (*multi-genre*) запрос средствами расширенной аналитики. Для этого анализируют запрос и разбивают его на несколько подзапросов, которые будут выполняться в виде

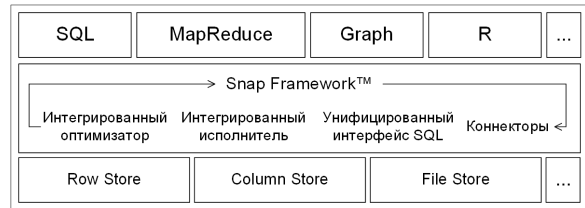


Рис. 7

конкретной рабочей нагрузки — специализированными программными обработчиками *Graph* и *MapReduce*. Интегрированный исполнитель организует выполнение запросов на узлах кластера и осуществляет контроль и управление над всеми обработчиками. Коннекторы обеспечивают доступ к источникам данных в социальных сетях, позволяют пользователям извлекать мультиструктурированные данные из *Hadoop*, *Teradata*, *Spark*, *Splunk* и любого другого источника данных. Подготовка данных и связанные с этим функции аналогичны *Teradata Aster Analytics*.

SNAP Framework[™] — это эксклюзивный продукт *Teradata*, основанный на стандартах, расширяемый и интегрируемый с существующими ИТ-инфраструктурами.

Платформа *Teradata Aster Discovery* (см. рис. 6) спроектирована для новых типов данных, объединяемых термином *BigData*. Это гарантирует использование всех доступных данных от различных источников (веб-журналы, сенсорные сети, социальные медиа, подробные данные, полученные от телекоммуникационных сетей и *Internet of Things*, астрономические и военные наблюдения, биологические системы, медицинские записи, фото- и видеоархивы), с возможностью размещения их безграничных объемов для предоставления необходимых сведений.

Синергетические мульти-жанровые средства аналитики используют множественные механизмы (*SQL*, *MapReduce*, статистические, анализ текста и *Graph*-анализ) для выявления новых идей. Расширенная библиотека предварительно построенных *SQL*, *SQL-MapReduce*, функций *Graph* позволяют начать сбор, подготовку, анализ и визуализацию данных одним

¹⁴ Аналитики *Gartner* употребляют иное выражение — мультисистемные (*multisystem*) запросы, — характеризуя *Teradata* в Отчете за 2017 г. [65].

действием. Обладая мощными аналитическими приложениями в сочетании с минимальными затратами времени на достижение результата и усилиям, затрачиваемым на это, *Teradata* предоставляет знания для современных компаний.

В завершение запущена *Teradata Cloud*, предоставляющая сложную аналитику как услугу с быстрой подготовкой и масштабируемостью в модели подписки. Помимо этого, БД *Teradata* и *Teradata Aster Discovery Platform* доступны как службы для хранилищ данных и аналитики обнаружения данных. Единая архитектура данных *UDA Teradata™* для всех типов данных с возможностью облачной доставки.

Teradata несомненно удалось создать платформу исследования мульти-структурированных данных для поддержания сложной и насыщенной данными аналитической нагрузки. Ключом к достижению успеха с коллекциями больших данных послужила возможность объединения нескольких типов данных и применение ряда аналитических методов к решению единой проблемы — выявлению данных, обуславливающих важнейшие неинтуитивные бизнес-идеи. Причем, всего этого можно достичь в одном решении путем единого запроса, объединяющего все шаги расширенных средств анализа, предназначенного для всех типов пользователей, который устраняет необходимость поддерживать разрозненные инструменты, нанимать дорогостоящих узкопрофильных специалистов с навыками работы с силовыми метаданными в «бункерах данных» — *Data Silos (Elevator)*. Термин *силос данных* [63] используется как иллюстрация факта отдельного существования и хранения фиксированных¹⁵ данных.

Например, геномные данные часто изолированы в силосах, что делает их недоступными для других исследователей. Силосы данных

— это результат технологии, препятствующей сотрудничеству. Они подвергаются критике не столько из-за препятствий производительности, но и потому, что силос отрицательно влияет на целостность данных — для одних и тех же данных могут существовать два или более внутренних хранилищ в крупных организациях. В этом случае их содержимое, вероятно, будет отличаться, создавая путаницу в отношении того, какой репозиторий представляет собой окончательную версию.

Teradata продолжает демонстрировать способность отслеживать рыночные тенденции, удовлетворяя требованиям клиентов. Например, в 2013 г. реализует инновационную технологию баз данных *Intelligent Memory Teradata (IMDBMS от Teradata)* — расширенное пространство памяти за пределами кеша, что значительно увеличивает производительность запросов и обеспечивает эффективную технологию хранения разнообразных данных в памяти. По сути, речь идет о использовании технологии *ИМС* [49, 50] и возможности выполнять гибридную транзакционную/аналитическую обработку данных в оперативной памяти *HTAP*. Направленность разработок *Teradata* отражает тенденции последнего десятилетия развития хранилищ данных и соответствует требованиям к ним и к средам управления данными для аналитики (*Data Warehouse and Data Management Solutions for Analytics — DW and DMSA*), изложенным *Gartner* в [23, 24, 26, 27, 30, 42]. Предложения *Teradata* предусматривают традиционные¹⁶ и логические (*LDW*) решения, которые *Teradata* называет *Unified Data Architecture (UDA)* [64].

Унифицированная архитектура данных *Teradata™ (UDA)* — это доступное и лучшее в своем классе аналитическое решение. Интегрируя хранилище данных *Teradata*, платформу для обнаружения *Aster* и *Hadoop* с открытым исходным кодом в единую прозрачную структуру

¹⁵ Фиксированные данные (исправленные, постоянные, справочные или архивные) — это данные, которые не могут быть изменены. Примеры фиксированных данных содержат результаты завершённых исследований, медицинских записей и исторических данных [63].

¹⁶ Традиционные хранилища владеют историческими данными, поступающими из различных структурированных источников.

(см. рис. 6), *Teradata Unified Data Architecture*TM устраняет разрыв между бизнес-языком *SQL* и растущей популярностью *MapReduce*. Поскольку в таком решении идея экосистемного подхода, использующего *Hadoop* наряду с реляционными и другими средами для охвата разного типа данных, задача сводится к связыванию узловых точек, хранящихся в разных средах (рис. 8). Результат — унифицированная, высокопроизводительная аналитическая среда с привлечением преимуществ специализированных процессоров обработки, — среда, обеспечивающая бесшовный межплатформенный доступ к данным разного рода и аналитическим механизмам обработки (рис. 9).

В унифицированной архитектуре данных *Teradata (UDA)* существует три основные платформы: хранилище данных (*Teradata Integrated Data Warehouse*), платформа обнаружения (*Teradata Aster Discovery*) и платформа данных (*Data Platform*). Большие стрелки (см. рис. 8) обозначают путь данных, проходящих между этими системами. Например, если нужно интегрированное аппаратное и программное решение (*Database Appliance*), оптимизированное специально для больших данных, *Data Platform* превращается в *Teradata® Appliance* для *Hadoop*, оставаясь в существующей инфраструктуре. В качестве решения для сбора данных, требующих масштабирования, *Teradata* запускает программное обеспечение *Hadoop*. Имеется возможность выбирать, какой дистрибутив *Hadoop (Hortonworks, Cloudera или MapR)* наилучшим образом соответствует потребностям, обеспечивая при этом интеграционные возможности и преимущества обслуживания и поддержки *Teradata*.

VIRTUAL QUERY — QueryGrid является центральным звеном унифицированной, высокопроизводительной аналитической среды (см. рис. 9). Экосистема *QueryGrid* обеспечивает беспрепятственный, высокопроизводительный доступ к данным, обработку и перемещение по системам в гетерогенных аналитических средах [65].

Концепт логических хранилищ данных (*LDW*) — единое представление данных, без их перемещения, — содержит, как компонент

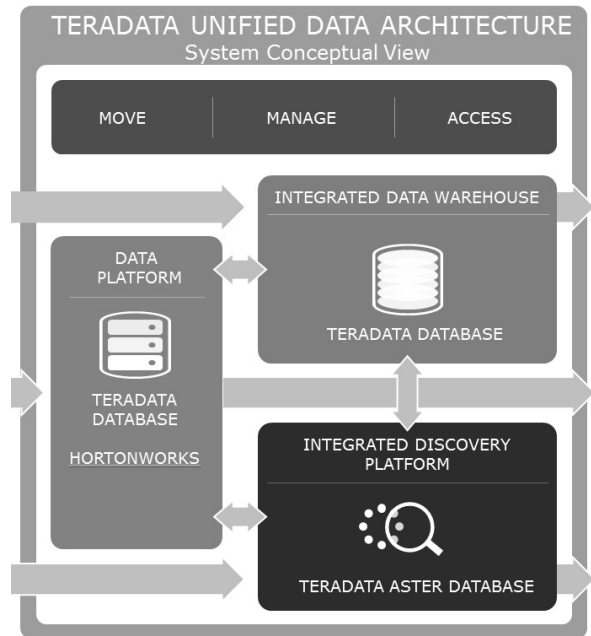


Рис. 8

LDW-структуры, виртуализацию данных,¹⁷ т.е. абстрагирование данных от конкретной формы их хранения. *LDW* должны поддерживать концепцию рендеринга памяти или указанных типов ресурсов данных, которые непосредственно считывают исходные системы [27]. В результате обеспечивается доступ ко всем видам информационных ресурсов, что является значительным шагом в интеграции информации. Логическое хранилище данных предоставляет платформу информационных услуг для приложений.

*Teradata QueryGrid*TM — действительно форма виртуализации данных, однако она не страдает от обычных ограничений виртуализации, таких как низкая производительность, перегруженные сети и проблемы безопасно-

¹⁷ Виртуализация данных — процесс предоставления данных посредством интерфейса, скрывающего все технические аспекты хранения данных (способ хранения, местоположение, структура, язык доступа). Логически местом для виртуализации данных служит дополнительный промежуточный уровень, изолирующий физическое хранение данных от приложений.

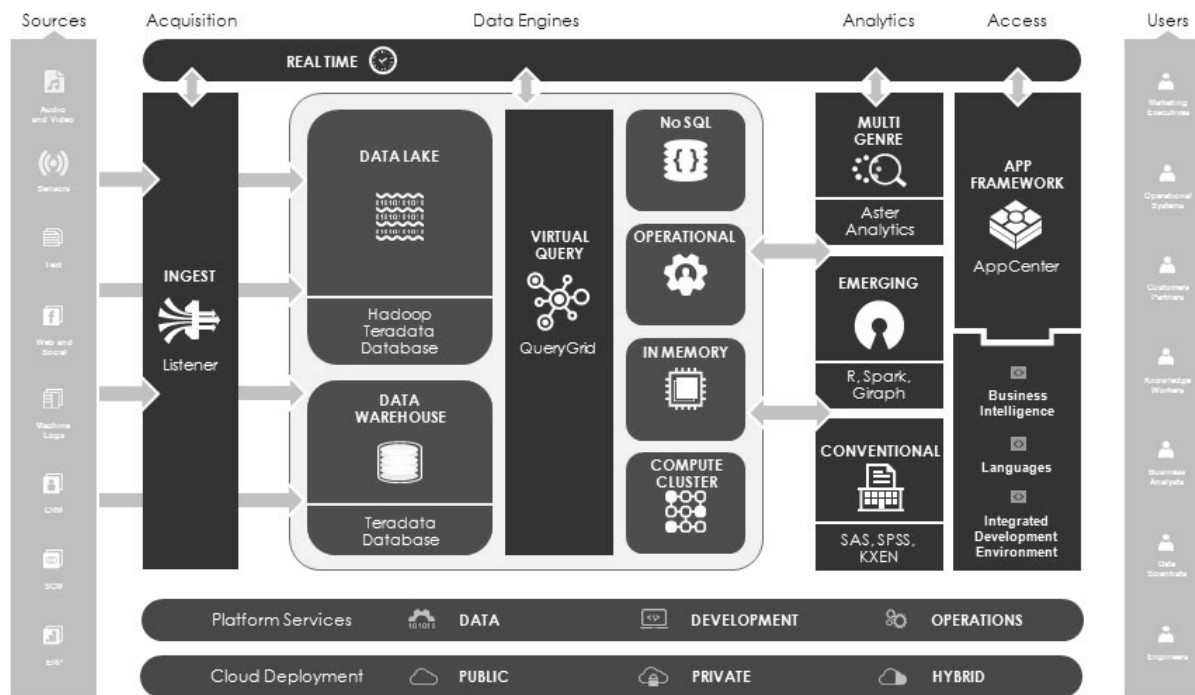


Рис. 9

сти. Исторически, средства виртуализации данных не имеют возможности перемещать данные между системами параллельно. Такие инструменты отправляют запрос в удаленную базу данных, и они возвращаются последовательно через *Ethernet*. *QueryGrid* построена для одновременного подключения к системам обмена данными через множество сетевых подключений одновременно. Параллельная обработка обеими системами делает обмен быстрым — со скоростью до *10TB/second* при использовании межсоединений *BYNET V5* на *InfiniBand*.

Кластер *Hadoop* будет иметь различное количество серверов в сравнении с *MPP*-системами *Teradata* или *Aster*. Эти системы запускают параллельный обмен данными путем сопоставления единиц параллелизма между двумя системами, т.е. все параллельные рабочие группы *Teradata*, называемые *AMP*, подключаются к рабочему узлу *Hadoop* для максимальной пропускной способности. В любое время, когда конфигурация изменяется, рабочие изменения совпадают [65].

*Teradata QueryGrid*TM [66] обеспечивает обработку запросов через источники данных, такие как БД *Teradata* и система *Hadoop*, на которой установлена *Teradata Distribution Presto*¹⁸. Коннекторы преобразуют язык запросов и данных, чтобы их можно было обменивать между различными типами систем, контролируют выполнение запросов в целевых системах и возвращают результаты в инициирующие системы. Коннекторы работают на одном или нескольких узлах в системе и относятся к типу системы, например, *Teradata* или *Presto*-скон-

¹⁸ *Presto* — распределенный механизм запросов с использованием стандартного *ANSI SQL* с открытым исходным кодом, предназначенный для запуска интерактивных аналитических запросов к различным источникам данных размером от гигабайт до петабайт. Его архитектура основана на скорости, что позволяет *Presto* поддерживать большое количество параллельных интерактивных запросов в среде больших наборов данных; позволяет получать доступ к данным в *Hadoop*, *Amazon S3*, *Apache Cassandra*TM, реляционных БД, и хорошо сочетается с единой архитектурой *Teradata Unified Data Architecture*TM.

фигурированный *Hadoop*. Реляционные таблицы в *Teradata* [65] могут просто комбинироваться с таблицами или плоскими файлами, найденными по требованию в *Hadoop*. Для этого не требуется извлекать данные из одной системы и загружать их на другую, и не нужно заботиться о том, где они находятся.

Teradata QueryGrid[™] обеспечивает высокую производительность, сокращая время обработки с часов до минут, вследствие автоматизированного и оптимизированного распределения работ посредством обработки «push down» на разных платформах, прозрачной автоматизации аналитической обработки, минимального и двунаправленного перемещения данных между системами, сведения к минимуму их дублирования, уменьшения нагрузки на сеть.

Коннекторы, обеспечивающие соединения, работают между *Teradata Database* и *Hadoop*, а также *Aster Databases* и *Hadoop*. *Teradata QueryGrid* также связывает хранилище данных *Teradata* с базами данных *Oracle*. В ближайшем будущем он будет распространяться на все комбинации серверов *UDA*, таких как *Teradata*, *Aster*, *Aster* на *Aster*, *Teradata* на *Teradata* и др. [65].

Новым в вертикали *Acquisition* (получение или сбор данных, см. рис. 9) является также модуль переноса и распределения чрезвычайно быстрых потоков данных в аналитические экосистемы — *INGEST & Listener*¹⁹. Модуль позволяет получать в реальном времени (*in real-time*) потоки данных большого объема (*Big Data Streaming*) и сохранять данные из этих потоков в *Teradata Aster* или *Hadoop* для обработки.

Появление *Internet of Things (IoT)* и постоянный рост потоковых данных создают новые проблемы при развертывании простых, надежных и масштабируемых систем сбора данных в режиме *real-time*. Особенно тех, которые нуждаются в самообслуживании для быстрой настройки новых данных, осуществляемой

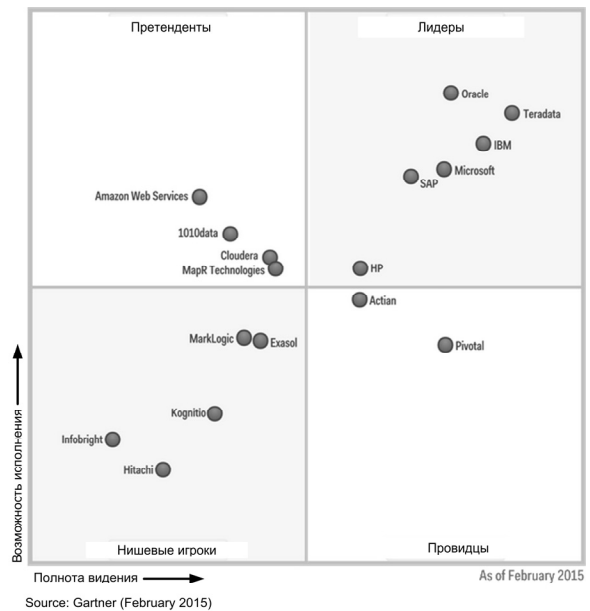


Рис. 10

посредством интеллектуального программного обеспечения *Teradata Listener*, осуществляющего в режиме *real-time* захват и настройку большого объема непрерывных потоков данных.

Платформа потоковых данных предназначена для сбора и управления большими объемами потоков в (*in real-time*) из любого источника, такого как веб-сайты, социальные каналы и пр. Позволяет обслуживать до 100 потоков, надежно передает данные без потерь и работает как программное решение только в центре обработки данных или в облаке. В целом модуль *INGEST & Listener* решает потребности потоковой передачи *BigData*, унифицируя большой процесс приема данных, непрерывно захватывая несколько потоков данных большого объема из разных источников и сохраняя их в одном или нескольких хранилищах данных.

Оценка Gartner. *Teradata* позиционируется как лидер в течение 15 лет подряд в *Magic Quadrant Gartner* (см. рис. 2) 2014 г. для систем управления хранилищами данных. В отчете *Gartner* следуют три аналитические оценки *Teradata* за 2014 г. [42]. *Первая* — лидерство в области аналитики, *вторая* — признание архитектуры *Tera-*

¹⁹ *Teradata Listener*[™] — <https://www.teradata.com/products-and-services/listener>

data® Unified Data Architecture™, и третья — лидерство в области использования *Hadoop* для больших объемов данных. Начиная с 2015 г., хранилища данных расширились для обращения к нескольким типам данных, механизмам обработки и репозиториям. Плотная группа лидеров (см. рис. 2) имеет более значительное расщепление и представлена в виде литеры «У» [23] в *Magic Quadrant* для хранилищ данных и сред управления данными для аналитики. Вершину правой ветви завершает *Teradata* (рис. 10).

Gartner, характеризуя рынок, отмечает, что организации теперь нуждаются в решениях, способных управлять и обрабатывать, в сочетании с данными из традиционных внутренних источников, внешние данные различных типов в разных форматах. Последние, например, могут включать в себя даже данные взаимодействия и наблюдения от интернет-датчиков *Things (IoT)*, а также нереляционные данные, такие как текст, изображения, звук и видео. Это создает новые требования на рынке хранилищ данных — при более широких решениях управления данными для аналитики *DMSA* с функциями и функциональностью, представляющими собой значительное увеличение существующих стратегий хранилищ данных [23, 30]. При использовании разнородных информационных ресурсов *Gartner* [23] выделяет четыре класса хранилищ: традиционные, операционные (*ODS*)²⁰, логические (*LDW*) и контекстно независимые хранилища²¹.

В дальнейшем эта тенденция продолжает расти: увеличивается спрос на решения, охватывающие различные типы данных и предложения вариантов стратегии хранилища. Рынок хранилищ данных продолжает развиваться и включает в себя более широкие решения для управления данными для аналитики под

координированным подходом, требующим различных типов интегрированных решений и уровня совместимых сервисов для управления и доставки данных. Менеджеры хранилищ данных, архитекторы решений для аналитики и другие, устанавливающие стратегии модернизации ИТ, должны учитывать это изменение направления и готовиться к гибридным технологическим платформам, расширяющим хранилище данных за пределами любой текущей практики [23, 24].²¹

В отчете *Gartner* 2017 г. [67] *Magic Quadrant* для решений управления данными для аналитики, отмечается, что *Teradata* построила платформу *DMSA*, учитывающую все варианты использования хранилищ данных: традиционное, операционное, логическое и контекстно независимое. *Teradata* даны наивысшие оценки продукта во всех четырех категориях. Эти оценки отражают логический подход к хранилищу, представленному унифицированной архитектурой данных *Teradata®*, включая *Teradata Aster Analytics* и *Hadoop*, и подчеркивают статус *LDW* как окончательного решения для баз данных для аналитики — во всех возможных случаях использования [68].

В определениях *Gartner* указано, что *DMSA* должна управлять хранением и доступом, процессами обработки и доставки данных, предназначенных для одного или нескольких из четырех основных случаев использования. *DMSA* не является конкретным классом или типом СУБД, она может состоять из множества различных технологий управления данными или их сочетания. Однако любое предложение или комбинация предложений должны в своей основе предоставлять доступ к

²⁰ *ODS* — операционные хранилища (оперативные склады данных) управляют структурированными данными, загружаемыми в систему непрерывно для вычисления аналитических функций приложений, встроенных в эти хранилища. Поддерживают отчетность и автоматизированные запросы для обеспечения текущего функционирования предприятия.

²¹ Контекстно независимые хранилища могут изменять схемы чтения данных, что позволяет получать новые информационные срезы, извлекая дополнительные сведения. Это достигается путем использования средств поиска, механизмов графов и других расширенных возможностей для извлечения новых информационных моделей. Обычно такие системы используются продвинутыми пользователями (исследователями данных), решающими нестандартные аналитические задачи.

данным под управлением средств открытого доступа через *API*; например, через *Open Database Connectivity (ODBC)*, *Java Database Connectivity (JDBC)*, *Object Linking and Embedding Database (OLEDB)* и др. Существует множество различных моделей доставки, таких как автономное программное обеспечение СУБД, сертифицированные конфигурации, платформа баз данных в качестве услуг (*dbPaaS*) и хранилищ данных [67, 68].

В *Teradata* эти решения содержат специальные стратегии оптимизации, предназначенные для поддержки аналитической обработки, учитывая (но не ограничиваясь), реляционную и нереляционную обработку (например, обработку графа) и машинное обучение или языки программирования, такие как *Python* или *R*. Данные не обязательно хранятся в реляционной структуре, можно использовать несколько моделей (реляционные, документ, ключ-значение, текст, граф, геопространственные и др.) [24, 42, 52, 68].

Предложения *Teradata* [67] включают в себя решение СУБД, устройства хранилищ данных и решения облачных хранилищ данных (все *MPP*) как на частном управляемом облаке, так и на инфраструктуре общественного облачного провайдера, такого как *Amazon Web Services (AWS)* и *Microsoft Azure*. Поддержка *LDW* поставляется с единой архитектурой данных (*UDA*). *Teradata QueryGrid* (часть *UDA*) обеспечивает поддержку мультисистемных (*multisystem*) запросов через *Teradata*, а также *open-source Presto*. *Teradata* также предлагает поддержку *Aster Analytics* и *Hadoop* через все три основные дистрибутивы, а также аналитические консалтинговые услуги.

Новая норма в управлении данными способствует объединению облачных и локальных сетей в гибридных конфигурациях. *Teradata IntelliCloud™* — это новое поколение защищенных облачных решений, предоставляющее данные и аналитическое программное обеспечение как услугу по модели *SaaS*. Услуги *IntelliCloud* (БД *Teradata*, *Aster® Analytics* или поддерживаемого программного обеспечения *Hadoop®* от *Cloudera* или *Horton-*

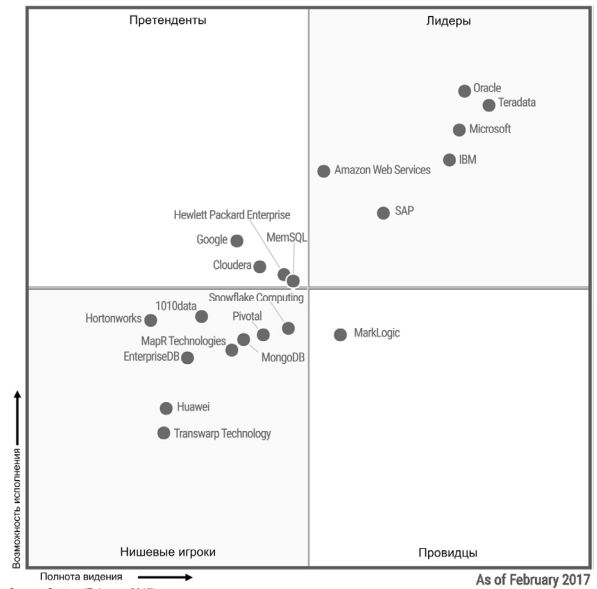


Рис. 11

works) доступны с новыми вариантами развертывания, охватывающие *Teradata IntelliFlex™* — флагманскую корпоративную платформу хранилищ данных предприятия, и *IntelliBase* в ЦОД *Teradata* и на инфраструктуре облака *AWS* и *Microsoft Azure*. Полную настройку инфраструктуры и управление программным обеспечением в обоих случаях осуществляет *Teradata* [52, 68].

База данных *Teradata* будет доступна на *Microsoft Azure Marketplace*, предоставляя клиентам возможность выбора полной базы данных *Teradata* по требованию и по принципу «оплата по мере необходимости». БД *Teradata* на *Azure* предложит полные функции и *MPP*-обработку, масштабируемую до 32 узлов. В ее реализации будет использована платформа *IntelliFlex* следующего поколения *Teradata*, предлагающая независимое масштабирование вычислительной мощности и емкости памяти, расширенные вычисления в памяти и сокращение времени простоя при масштабировании системы.

В отчете *Gartner 2017 г.* подчеркнуты облачные решения и возможности аналитической обработки, отходя от намерения сосредоточиться на реляционных БД с использованием *SQL*. Облако стало стандартом де-факто. Те,

кто не принимает его, обречены на устаревание. Следует обратить внимание на прыжок, который совершил *AWS* в квадрант «Лидеры» с позиции «Претенденты» (см. рис. 2, 10; рис. 11). Тенденция облака не только в секторе *DMSA*, она распространяется на всю ИТ-отрасль. Борьба между хранилищем данных и новыми моделями логических хранилищ — это разногласие между традиционным коммерческим программным обеспечением и новой экосистемой с открытым исходным кодом. Все поля магического квадранта переживают революцию, поскольку возникают бесчисленные стартапы. То же относится к анализу и обработке данных, поскольку этот элемент также перемещается в облако [68].

Компания *Teradata* — ведущая в области аналитических решений — объявила о начале работы первой в мире, всюду доступной *Teradata Everywhere*™, предоставляющей самую мощную аналитическую базу данных с широким спектром параллельных вычислений (*MPP*) для локальных сред, множества публичных облаков, управляемого частного облака. По мере того, как компании продвигаются к гибридным облачным архитектурам, *Teradata* лидирует в отрасли, предлагая клиентам выбор и обеспечение совместимости баз данных в режимах развертывания. *Teradata Everywhere* обеспечивает гибкость для реализации гибридной архитектуры с общей базой данных, позволяющей перемещать рабочие нагрузки между средами по мере развития потребностей бизнеса, поддерживая изменяющуюся стратегию развертывания компании и экономические потребности [9].

Teradata достигает наивысшей позиции в *Magic Quadrant* благодаря стремлению к совершенству в современных аналитических технологиях, своей прозорливости, гибкости в программно-технических решениях и полноте видения рыночной ситуации для *DMSA* [52, 68, 69].

Заключение

Подготовке этого раздела обзора предшествовал многочасовой период просмотра материа-

лов сайта *Quora*²² и сверки собственного видения предмета изложения с мнением разработчиков и богатым опытом пользователей *Teradata*. Это реляционная система параллельной обработки, в которой используется архитектура без общего доступа. В ее основе — технология, состоящая из оборудования, программного обеспечения, базы данных и консалтинга. Система перемещает данные в хранилище, где их можно вызвать и проанализировать.²²

Создавав свое решение (*Database Appliance*) между специализированными аппаратными средствами и программным обеспечением, *Teradata* успешна в нише *Data Warehouse* на протяжении длительного времени, достигнув производительности в очень больших базах данных в задачах аналитики при принятии стратегических решений. Но *Teradata Database Appliance* — прибор, а это значит, что пользователь либо делает это «методом *teradata*», либо нет. Невозможны настройки оборудования на проблемное пространство. Когда требуется масштабировать аппаратное обеспечение, следует использовать весь прибор. Нет никаких вариантов использования оптимизации облачного или эластичного стиля. Также отмечено, что *Teradata* имеет незрелый уровень абстракции данных. Незначительные изменения в программировании возможны в сравнении с другими языками РСУБД.

Teradata своевременно отреагировала на насущную необходимость анализа *BigData* и в первую очередь данных новых форматов медиаисточников, *IoT* и др. Первым решением был *Hadoop for the Enterprise* — гибкий набор аппаратных средств, программного обеспечения и сервисов для интеграции *Hadoop* в среду *Teradata*. Затем — платформа для обнаружения дан-

²² *Quora* — социальный сервис обмена знаниями, своеобразный онлайн-рынок вопросов и ответов, основанный в июне 2009 г. Фишкой сервиса и отличием от форумов являются профессиональные ответы на вопросы. *Quora* имеет специальный штат сотрудников, которые и занимаются поиском ответов на заданную тематику.

ных *Teradata Aster Discovery*. Разработаны и использованы аналитические функции *SQL*, *SQL-MapReduce*®, *Graph*, функции временных рядов, статистических методов, аналитики текста и многое другое для исследования *BigData*, обеспечивается доступ к мультиструктурированным данным в *Apache™ Hadoop™*, *Teradata Data Warehouse* и других реляционных СУБД (*RDBMS*).

Высокопроизводительный доступ к данным, обработку и виртуальную доставку к системам в гетерогенных аналитических средах обеспечивает экосистема *Teradata Query-*

Grid™ — своеобразная матрица, использующая параллельное перемещение данных между системами обмена. Идея экосистемного подхода, применяющего *Hadoop* наряду с реляционными и другими средами для охвата разного типа данных, сводится к задаче связывания узловых информационных точек в разных средах. Принятая унифицированная архитектура данных *Teradata™ (UDA)* не противоречит единному представлению данных без их перемещения, — концепту логических хранилищ данных.

REFERENCES

52. *Teradata*, <https://www.teradata.com/>
53. *Dewitt David, Gray Jim*. Параллельные системы баз данных: будущее высоко эффективных систем баз данных, 02 Apr. 1995, https://www.osp.ru/news/articles/1995/0402/13031420#part_2_1
54. Блог компании Teradata: Teradata, СУБД, параллельная от рождения, 3 дек. 2012, <http://habrahabr.ru/company/teradata/blog/160821/>
55. Блог компании Teradata: Статистика в СУБД Teradata, 1 февр. 2013, <http://habrahabr.ru/company/teradata/blog/167801/>
56. Физический дизайн структур хранения в СУБД Teradata, 16 янв. 2014, <http://habrahabr.ru/company/teradata/blog/170321/>
57. Блог компании Teradata: Дополнительные техники физического моделирования в Teradata, 17 янв 2014, <http://habrahabr.ru/company/teradata/blog/209216/>
58. Блог компании Teradata: Скорость или объем? Автоматизация управления системами хранения с разнородными характеристиками, 11 янв. 2013, <http://habrahabr.ru/company/teradata/blog/165463/>
59. *MapReduce* and *Teradata Aster SQL-MapReduce*®, <https://www.teradata.com/products-and-services/Teradata-Aster/teradata-aster-sql-mapreduce>
60. *Teradata Aster Discovery Portfolio*, http://assets.teradata.com/resourceCenter/downloads/Brochures/Teradata_Aster_Discovery_Portfolio_EB7509.pdf
61. *Teradata Aster Analytics*, http://assets.teradata.com/resourceCenter/downloads/Brochures/Teradata_Aster_Discovery_Platform_EB7573.pdf?processed=1
62. *Dean J*. MapReduce: Simplified Data Processing on Large Clusters, <http://research.google.com/archive/mapreduce.html>
63. What is data silo? / Definition..., <http://searchcloudapplications.techtarget.com/definition/data-silo>
64. *Teradata Unified Data Architecture™*, <https://www.teradata.com/Solutions-and-Industries/unified-data-architecture> или <https://www.teradata.com/Resources/White-Papers/Teradata-Unified-Data-Architecture-Integrate>
65. *Take a Giant Step with Teradata QueryGrid*, <http://blogs.teradata.com/data-points/take-a-giant-step-with-teradata-querygrid/>
66. *Teradata QueryGrid User Guide*, *Teradata Documentation*, <https://www.info.teradata.com/download.cfm?ItemID=1007085>
67. *Magic Quadrant for Data Management Solutions for Analytics* / R. Edjlali, Adam M. Ronthal, R. Greenwald et al, 28 Febr. 2017, <https://www.gartner.com/doc/reprints?id=1-3TZLPYX&ct=170221&st=sb>
68. *Teradata Achieves Highest Position for Completeness of Vision in Data Management Solutions for Analytics Magic Quadrant*, 28 Febr. 2017, <http://www.teradata.ru/News-Releases/2017/Teradata-Achieves-Highest-Position-for-Comple>
69. *Teradata Announces the World's Most Powerful Analytic Database, Available Everywhere*, 12 сент. 2016, <http://www.teradata.ru/News-Releases/2016/Teradata-Announces-the-World%E2%80%99s-Most-Powerful>

Поступила 16.01.2018

A.A. Oursatyev, PhD in Techn. Sciences, Leading Research Associate, International Research and Training Centre of Information Technologies and Systems of the NAS and MES of Ukraine, Glushkov ave., 40, Kyiv, 03187, Ukraine, aleksei@irtc.org.ua

BIG DATA. ANALYTICAL DATABASES AND DATA WAREHOUSE: TERADATA

Introduction. The article is a continuation of the Big Data and tools study, which is being transformed into technology of the new generation and architecture of the *BD* platforms and storage for the intelligent output. In this part the review of *DB Teradata* is presented. The main attention is paid to the issues of changing the infrastructure, the tool environment and the platform for identifying the necessary information and new knowledge from the Big Data, the initial information about the product is given in the product general description.

Purpose. The purpose is to consider and evaluate the application effectiveness of the infrastructure solutions for new developments in the Big Data study, to identify new knowledge, the implicit connections and in-depth understanding, insight into phenomena and processes.

Methods. The informational and analytical methods and technologies for data processing, the methods for data assessment and forecasting, taking into account the development of the most important areas of the informatics and information technology.

Results. *Teradata* is a relational system of the parallel processing, in which the architecture is used without general access. It is based on technology, consisting of equipment, software, databases and consulting. The system moves data to the storage area where they can be called up and analyzed.

Having created a solution (*Database Appliance*) between specialized hardware and software, *Teradata* is successful in the Data Warehouse for a long time, achieving performance in very large databases in the analytical tasks in making strategic decisions. But the *Teradata Database Appliance* is an instrument with the consequences as follow: there is no possibility of setting up the equipment for the problem space. When you need to scale hardware in one direction or another, you should use the entire device. There are no options for using cloud or elastic style optimization. It is also noted that *Teradata* has an immature level of data abstraction. Small changes in programming can be made in comparison with other languages *RDBMS*.

Teradata responded in a timely manner to the urgent need for *BigData* analysis and first of all these new formats of media sources. The first solution was *Hadoop* for the Enterprise, a flexible set of hardware, software and services for integrating *Hadoop* into the *Teradata* environment. Then it was a platform for detecting data *Teradata Aster Discovery*. The analytical functions *SQL*, *SQL-MapReduce*®, *Graph*, time series functions, statistical methods, text analytics and much more for *BigData* study are developed and used, the access to multi-structured data is provided in *Apache*™ *Hadoop*™, *Teradata Data Warehouse* and other relational database management system (*RDBMS*).

Teradata QueryGrid™ ecosystem provides high-performance data access, processing and virtual delivery to systems in heterogeneous analytical environment. This is a kind of matrix that uses parallel data transfer between the exchange objects. The idea of an ecosystem approach that uses *Hadoop* along with the relational and other environments to cover different types of data is reduced to the task of linking nodal information points stored in the different environments. The accepted unified data architecture *Teradata*™ (*UDA*) does not contradict the unified representation of data, without their movement, — the concept of the logical data stores.

Conclusion. According to Gartner's analysts in 2014 *Teradata* received three awards: leadership in the field of analytics, recognition of the *Teradata*® *Unified Data Architecture*™ architecture, and leadership in the use of *Hadoop* for the large amounts of data. Since 2015, data warehouses have expanded due to the several types of data access, processing mechanisms and repositories. In this regard, in the Gartner report 2017, it is noted that *Teradata* has built a data management platform that takes into account all the uses of data warehouses: the traditional, operational, logical and context-independent. This reflects the approach to the storage represented by the unified *Teradata*® *UDA*™ data architecture.

Teradata offers also provide solutions for cloud data warehouses both on a private managed cloud and on the infrastructure of a public cloud provider. This facilitates the integration of cloud and local networks in hybrid configurations. *Teradata IntelliCloud*™ is the next generation of secure cloud solutions that provides data and analytics software as a *SaaS* service.

Also, *Teradata* provides the solutions for cloud data warehouses both on a private managed cloud and on the infrastructure of a public cloud provider. This facilitates the integration of cloud and local networks in hybrid configurations. *Teradata IntelliCloud*™ is the next generation of secure cloud solutions that provides data and analytics software as a *SaaS* model.

Keywords: *MPP*, *Logical Data Warehouse LDW*, *Warehouse Appliance*, *SN (Shared Nothing)*, *Teradata Aster Discovery*, *Teradata IntelliCloud*™, *SaaS*, *Teradata QueryGrid*™, *Unified Data Architecture Teradata*™ (*UDA*).

О.А. Урсаев, канд. техн. наук, Міжнародний науково-навчальний центр інформаційних технологій та систем НАН України та МОН України, просп. Глушкова, 40, Київ 03187, Україна, aleksei@irtc.org.ua

ВЕЛИКІ ДАНІ. АНАЛІТИЧНІ БАЗИ ДАНИХ І СХОВИЩА: TERADATA

Вступ. Стаття є продовженням досліджень Великих Даних та інструментарію, що трансформується в нове покоління технологій і архітектури платформ БД та сховищ для інтелектуального виводу. У даній частині огляду подано *DB Teradata*. Основну увагу приділено питанням зміни інфраструктури, інструментального середовища і платформи для виявлення необхідної інформації та нових знань з Великих Даних, а початкові відомості про продукт наведено в загальній характеристиці виробу.

Мета. Розглянути та оцінити ефективність застосування інфраструктурних рішень нових розробок в дослідженнях Великих Даних для виявлення нових знань, неявних зв'язків і поглибленого розуміння, проникнення в сутність явищ і процесів.

Методи. Інформаційно-аналітичні методи і технології обробки даних, методи оцінки та прогнозування даних, з урахуванням розвитку найважливіших галузей інформатики та інформаційних технологій.

Результати. *Teradata* — це реляційна система паралельної обробки з використанням архітектури без загального доступу. В її основі технологія, що складається з обладнання, програмного забезпечення (ПЗ), бази даних та консалтингу. Система переміщує дані в сховище, де їх можна викликати та проаналізувати. Створюючи своє рішення — *Database Appliance* — між спеціалізованими апаратними засобами і ПЗ, *Teradata* успішна в ніші *Data Warehouse* протягом тривалого часу, досягнувши продуктивності в дуже великих базах даних в задачах аналітики при прийнятті стратегічних рішень. Але *Teradata Database Appliance* — це прилад з наслідками: немає можливості налаштування обладнання на проблемний простір. Немає жодних варіантів використання оптимізації хмарного або еластичного стилю. Так само відзначається, що *Teradata* має дуже незрілий рівень абстракції даних. Незначні зміни в програмуванні можливі в порівнянні з іншими мовами РСУБД.

Teradata своєчасно відреагувала на нагальну необхідність аналізу *BigData* і в першу чергу даних нових форматів медіаджерел. Першим рішенням був *Hadoop for the Enterprise* — гнучкий набір апаратних засобів, програмного забезпечення і сервісів для інтеграції *Hadoop* в середу *Teradata*. Потім — платформа для виявлення даних *Teradata Aster Discovery*. Розроблено та використано аналітичні функції *SQL*, *SQL-MapReduce*®, *Graph*, функції часових рядів, статистичних методів, аналітики тексту, забезпечено доступ до мультиструктурованих даних в *Apache Hadoop*™, *Teradata Data Warehouse* та інших реляційних СУБД.

Високопродуктивний доступ до даних, обробку і віртуальну доставку до систем в гетерогенних аналітичних середовищах забезпечує екосистема *Teradata QueryGrid*™ — своєрідна матриця, яка використовує паралельне переміщення між системам обміну. Ідея екосистемного підходу, який застосовує *Hadoop* поряд з реляційними і іншими середовищами для охоплення різного типу даних, зводиться до задачі зв'язування вузлових інформаційних точок, що зберігаються у середовищах. Прийнята уніфікована архітектура даних *Teradata*™ (*UDA*) не суперечить єдиному поданню даних, без їх переміщення, — концепту *LDW*.

Висновки. На думку аналітиків *Gartner*, за 2014 р. *Teradata* отримала три оцінки: лідерство в галузі аналітики, визнання архітектури *Teradata*® *Unified Data Architecture*™, та лідерство у використанні *Hadoop* для великих обсягів даних. З 2015 р. сховища даних розширилися внаслідок звернення до декількох типів даних, механізмів обробки та репозиторіям. У цьому зв'язку у звіті *Gartner* 2017 р. відзначається, що *Teradata* побудувала платформу управління даними, що враховує всі варіанти застосування сховищ даних: традиційну, операційну, логічну та контекстно-незалежну. Це відображає підхід до сховища з уніфікованою архітектурою даних *Teradata*® *UDA*™.

Пропозиції *Teradata* також містять рішення хмарних сховищ даних як на приватній керованій хмарі, так і на інфраструктурі громадського хмарного провайдера. Це сприяє об'єднанню хмарних і локальних мереж в гібридних конфігураціях. *Teradata IntelliCloud*™ — це нове покоління захищених хмарних рішень, що надає дані і аналітичне програмне забезпечення як послугу за моделлю *SaaS*.

Ключові слова: архітектура *MPP*, логічні сховища даних *LDW*, платформа *Warehouse Appliance*, *SN (Shared Nothing)*-архітектура, *Teradata Aster Discovery* — платформа виявлення, *Teradata IntelliCloud*™ — дані та аналітика за моделлю *SaaS*, *Teradata QueryGrid*™, уніфікована архітектура даних *Teradata*™ (*UDA*).