ИНФОРИНО 2020

Москва, Россия
14-17 апреля 2020 г.

Интеграция технологий Big Data и хранилищ данных

Авторы:

А.А. Еремеев, В.И. Еремеева

Национальный исследовательский университет «МЭИ», Москва

Докладчики:

Алексей Еремеев,

Валерия Еремеева

НИУ «МЭИ»

IEEE

МЭИ | НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

КРОК

Интеграция технологий Big Data и хранилищ данных

Большой интерес к технологиям класса BIG DATA связан с постоянным ростом данных, которыми приходится оперировать крупным компаниям. Актуальность темы статьи обусловлена сложностью и дороговизной обработки и извлечения пользы из накопленных данных.

Интеграция технологий Big Data и хранилищ данных

Термин “Big Data" характеризуется следующими признаками:

Volume – объем данных,

Velocity – необходимость обрабатывать информацию с большой скоростью,

Variety – многообразие и часто недостаточную структурированность данных.

Интеграция технологий Big Data и хранилищ данных

Технологию «Больших данных» следует рассмотреть совместно с технологией хранилищ данных. При рассмотрении основных технологических особенностей можно наблюдать как различия, так и области конвергенции. Даже при существовании различий обе технологии должны быть интегрированы, так как они нацелены на выполнение одной цели – поиск данных и поддержка принятия решений.

Интеграция технологий Big Data и хранилищ данных

Характеристика	Традиционная БД	База Больших Данных
Объем информации	От гигабайт(109 байт) до терабайт (1012 байт)	От петабайт(1015 байт) до эксабайт (1018 байт)
Способ хранения	Централизованный	Децентрализованный
Структурированность данных	Структурирована	Полуструктурирована и неструктурирована
Модель хранения и обработки данных	Вертикальная модель	Горизонтальная модель
Взаимосвязь данных	Сильная	Слабая

Интеграция технологий Big Data и хранилищ данных

Существующие решения:

- Oracle ;
- SAP Data Warehousing;
- Pentaho.

Упомянутые решения учитывают жизненный цикл данных, но больше ориентированы на обработку технологий, чем на интеграционные архитектуры и саму методологию.

Интеграция технологий Big Data и хранилищ данных

С точки зрения логической архитектуры, хранилища данных (ХД) и большие данные (Big Data) состоят из одинаковых компонентов:

источники данных;

процессы извлечения, преобразования и загрузки (ETL);

хранение, обработка и анализ.

Интеграция технологий Big Data и хранилищ данных

Источники данных:

- Структурированные данные
- Неструктурированные данные
  - повторяющиеся
  - неповторяющиеся

Интеграция технологий Big Data и хранилищ данных

При построении хранилищ данных необходимо учитывать ряд проблем, связанных с качеством данных, например, дублированные данные, возможную несогласованность данных, ненужные данные, создание новых переменных с использованием преобразований и т. д.

Интеграция технологий Big Data и хранилищ данных

В связи с необходимостью управления неструктурированными повторяющимися данными и неструктурированными неповторяющимися данными, поступающими из различных источников, предлагаются новые требования, среди которых можно выделить следующее:

- управление экспоненциальным ростом данных;
- частота поступления данных;
- долговечность, частота и возможность использования;
- интеграция данных.

Интеграция технологий Big Data и хранилищ данных

Big Data и новое поколение ХД не имеют предопределенных аналитических моделей и не полагаются на архитектуры клиент-сервер и должны поддерживать горизонтальное масштабирование. Ответом на новые потребности является использование обширной памяти, распараллеливание данных и обработка, которые так или иначе включены в базы данных Hadoop, MapReduce, NoSQL, хранение и обработка в памяти и технологии, дополняющие их.

Интеграция технологий Big Data и хранилищ данных

Предлагаемая многослойная архитектурная модель для Big Data состоит из трех уровней:

1) загрузка данных;
2) обработка и хранение данных;
3) анализ данных.

Интеграция технологий Big Data и хранилищ данных

Уровень загрузки данных специализируется на хранении в соответствии с типом данных.

Структурированные данные подвержены предварительной обработке и хранению согласно структурам и стандартным алгоритмам. Неструктурированные (повторяющиеся или неповторяющиеся) данные должны храниться в сыром виде и без контекста, для этой цели можно использовать распределенные базы данных NoSql.

Интеграция технологий Big Data и хранилищ данных

Структурированные данные агрегируются по заранее определенной модели. В то время как неструктурированные (повторяющиеся или неповторяющиеся) данные требуют процесса категоризации и фильтрации для хранения.

Интеграция технологий Big Data и хранилищ данных

Данная работа может быть полезна разработчикам программного обеспечения, которые должны формулировать и управлять Big Data или ХД проектами на уровне предприятия, так как предлагаемые решения должны быть поддержаны в масштабируемой модели архитектуры, которая позволяет обрабатывать новые источники и типы данных в систематическом порядке.

ИНФОРИНО 2020

Москва, Россия
14-17 апреля 2020 г.

Спасибо за внимание!

Контакты докладчика:

Алексей Еремеев
НИУ «МЭИ»

Валерия Еремеева
НИУ «МЭИ»

YeremeevAA@mpei.ru,
YeremeevaVI@mpei.ru

IEEE

МЭИ | НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

КРОК