МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

Инструменты для хранения и обработки больших данных

Big Data Storage and Processing Tools

МОСКОВСКИЙ
ГОРОДСКОЙ
УНИВЕРСИТЕТ
МГПУ

Современный анализ данных

Современный анализ данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Все данные, связанные с большими данными, генерируются ежесекундно и доступны из различных источников; именно разнообразие структур данных усложняет хранение и анализ.

В результате этого большие данные классифицируются по **3** основным операциям:

1. **хранение,**
2. **аналитика,**
3. **Интеграция.**

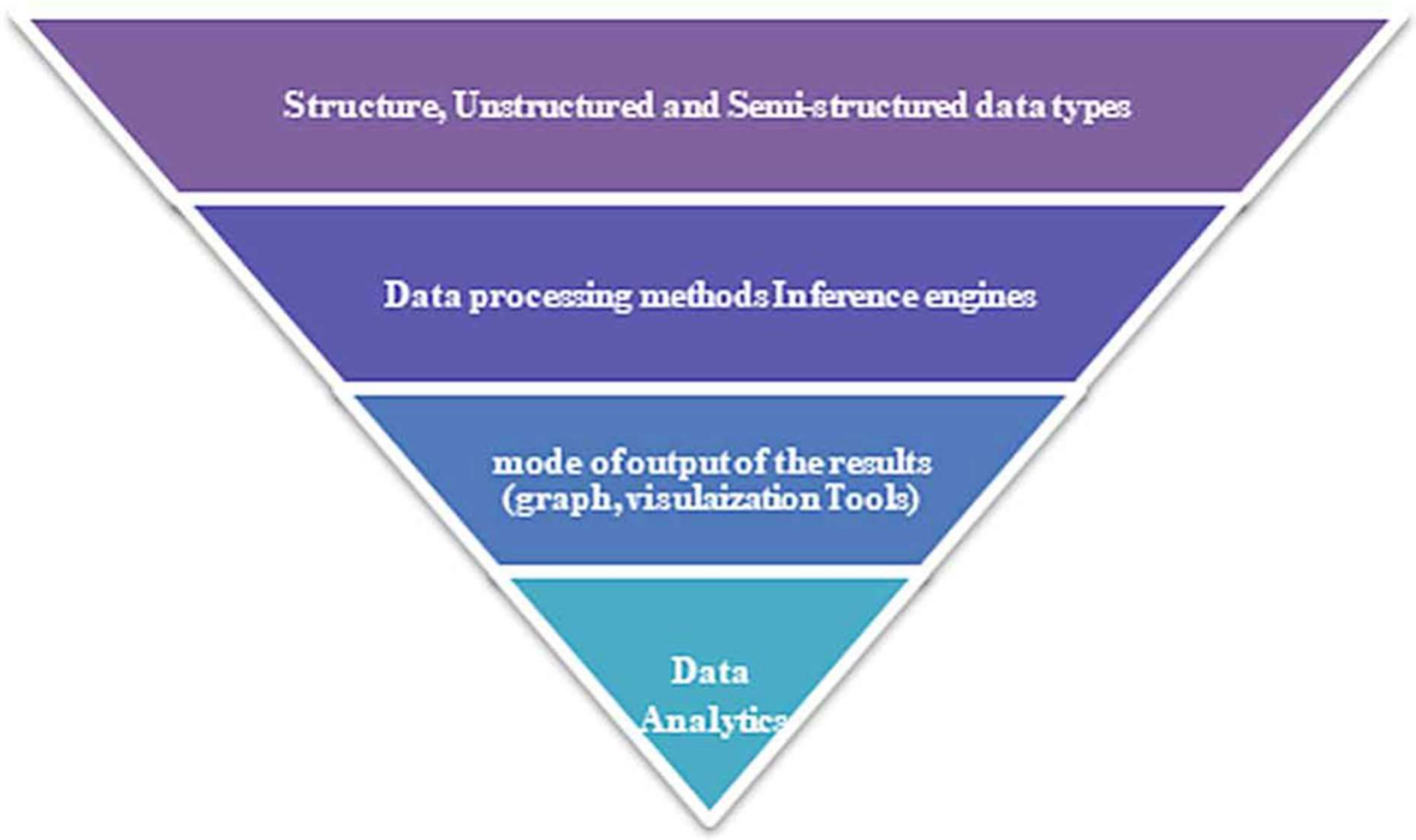
Потребность в аналитике больших данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Анализ больших данных решил многие проблемы, связанные с исследованиями в реальном времени, которые, в свою очередь, преобразовываются в другие форматы данных внешних информационных систем , таких как, например, графическая структура.

Потребность в аналитике больших данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ



Обработка больших данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ
МГПУ

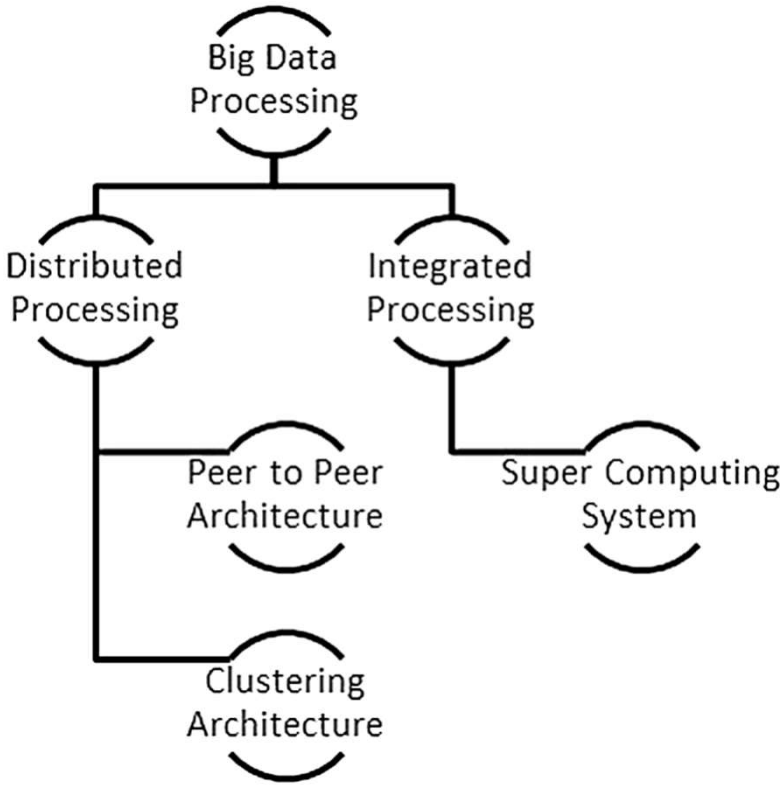
1. Интегральная обработка.
2. Распределенная обработка.

Обработка больших данных

1. Интегральная обработка.

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Данные собираются и хранятся в одном общем месте, где хранение и обработка выполняются на одном вычислительном узле(ноде). Этот интегрированный узел обработки включает в себя мощный процессор и память. Этот тип структуры обработки определяется ограниченной масштабируемостью, где хранение и обработка производится на одной машине. Такие системы называют - суперкомпьютерной системой.

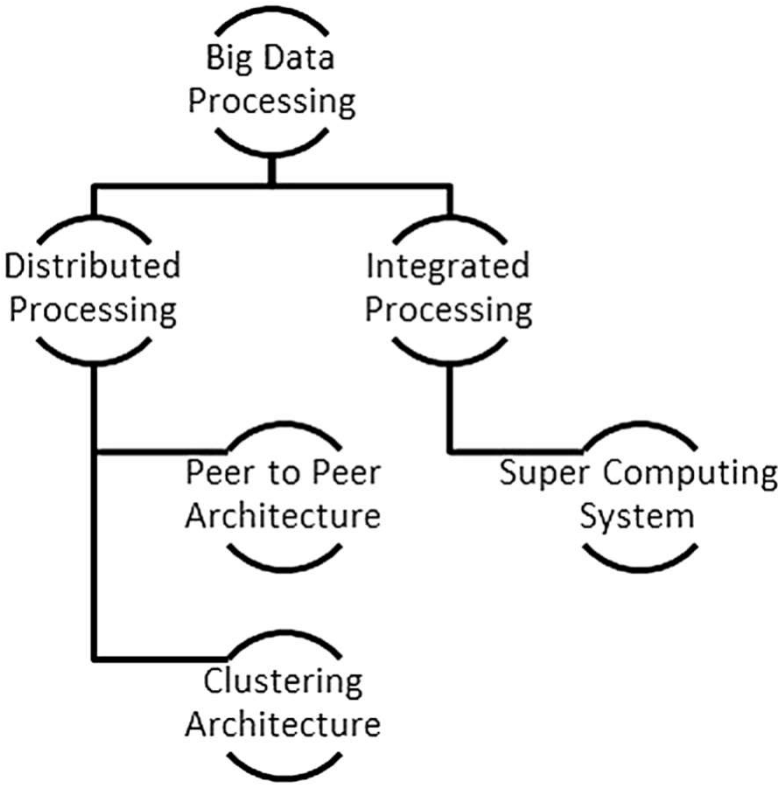


Обработка больших данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

2. Распределенная обработка.

Сбор, хранение и обработка выполняются на нескольких вычислительных узлах(нодах). Одноранговая архитектура и архитектура кластеризации являются ярким примером распределенной обработки.



Этапы анализа больших данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Большие данные используют очень большой набор данных, который не может быть обработан классическими инструментами и методами. Рассмотрим основные действия, которые необходимы для предобработки неструктурированных больших данных.

Этапы анализа больших данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

1. Сбор данных (Data Collection).

Получение необработанных данных из ресурсов генерации данных в режиме реального времени, которые затем сохраняются на устройстве хранения. Процесс сбора данных осуществляется предельно аккуратно, сбой в этом процессе приводит к получению неточных и не полных данных. Для сбора данных используются такие инструменты, как Chekwa, WebCrawler, pig и flume.

Этапы анализа больших данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

2. Разделение данных (Data Partition).

Поскольку большие данные трудно обрабатывать, существуют различные методы разделения, такие как:

- **классификация тегов данных,**
- **инструменты кластеризации,**
- **майнинг,**
- **Mahout,**
- **масштабируемый алгоритм ближайшего соседа,**
- **обработка шаблонов.**

Этапы анализа больших данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

3. Координация данных (Data Coordination).

Координация означает движение данных к любому хранилищу данных или к хранилищу данных на основе СУБД. То есть преобразование данных из одного формата в другой.

Sqoop — это технология больших данных, используемая для обмена данными из реляционной базы данных.

Flume — это технология, используемая в больших данных, предназначенная для эффективного хранения и управления огромным объемом данных из одной сети в другую.

Zookeeper обеспечивает синхронизацию данных, который используется для координации данных с использованием языка программирования информации о конфигурации, такого как java, python.

Этапы анализа больших данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

4. Преобразование данных(Data Transformation).

Преобразование одного формата данных из источников в другой формат относится к преобразованию данных. Инструменты миграции данных помогают преобразовать реляционную базу данных в репозиторий Hadoop.

Этапы анализа больших данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

5. Хранилище данных (Data Storage).

Хранилище данных должно эффективно выполнять поиск, обработку и сбор данных. Обработка различных типов данных также играет роль в хранении данных. Используемые инструменты: **HBase**, **NoSQL**, кластер **HDFS** и **GFS**.

Этапы анализа больших данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

6. Обработка данных (Data Processing):

До сих пор нет единого определения и инструмента разработки для обработки больших данных.

Hadoop, NoSQL, apache-spark и т. д. помогают обрабатывать структурированные и неструктурированные данные в различных форматах.

QlikView — это идеальный пример обработки данных в памяти для больших данных, который предоставляет расширенные отчеты.

Infinispan — это масштабируемая и доступная обработка данных грид-платформы.

Этапы анализа больших данных

7. Извлечение данных (Extract Data).

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Извлечение необходимых файлов или данных из базы данных с получением предпочтительного вывода в различных отчетах о результатах, таких как визуализация, интеграция и отчетность.

Два метода, используемые для извлечения данных. Инструменты запроса данных с использованием языка запросов, такого как Hive, помогают в извлечении данных. Поиск больших данных — использование параллельной или распределенной обработки с кластеризацией выборки данных.

Этапы анализа больших данных

8. Анализ данных(Data Analysis).

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Анализ данных определяется как предобработка, исследование или анализ и проверка данных, а затем моделирование в соответствии с целью, путем создания полезных данных.

Rapid Miner — это программное обеспечение с открытым исходным кодом, в котором текст добывается для использования данных для прогнозного анализа.

Pentaho - это программное обеспечение для бизнес-аналитики, где есть видео; данные OLAP, Сервис, ETL используются.

Talend и Spago BI – это инструменты, которые используются во многих управленческих организациях.

Weka — это инструмент машинного обучения, в котором для анализа данных реализован алгоритм интеллектуального анализа данных.

Этапы анализа больших данных

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

9. Визуализация данных (Data Visualisation)

DIVE и **Orange** — это средства визуализации больших данных, используемые для форматирования огромных данных в структурированный формат.

Инструменты, используемые для анализа данных: с точки зрения управления хранилищем

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Рассмотрим основные инструменты, которые в настоящее время используются для хранения, управления и анализа данных при формировании больших данных.

Инструменты, используемые для анализа данных: с точки зрения управления хранилищем

МОСКОВСКИЙ ГОРОДСКОЙ УНИВЕРСИТЕТ МГПУ

Hadoop (HDFS): расшифровывается как распределенная файловая система, которая, по прогнозам, будет работать на стандартном оборудовании изолированного узла сети(ноды).

Программное обеспечение Apache

Инструменты Apache	Краткое описание	Категория	Файл данных	Язык
Airavata	Airavata is a software framework, mini-service architecture used to implement and manage the flow of work and reckoning job. They use distributed computing resources.	Cloud, Big data and network-server	DOAP RDF (json)	Java
Ambari	Software Framework to process Hadoop cluster and other data processing domains.	Big data	DOAP RDF (json)	Java, Python and JavaScript
Apex	Batch processing Search engine	Big-Data	DOAP RDF (json)	Java
Avro	Data Serialization System	Library, Big Data	DOAP RDF (json)	C, C++, C#, Java, PHP, Python, Ruby
Beam	Programming model runs with data processing pipelines	Big-Data	DOAP RDF (JSON)	Java, Python

Программное обеспечение Apache

Bigtop	Community- driven BigData management platform	Big-Data	DOAP RDF (json)	Java
BookKeeper	Authentic Log service	Big-data	DOAP RDF (json)	Java
Calcite	Dynamic data Management Framework	Big-Data, Hadoop, SQL	DOAP RDF (JSON)	Java
Couch DB	NoSQL- Database using JSON and MapReduce and HTTP	Database, Big-data	DOAP RDF (json)	JavaScript, Erlang, C++,C
Crunch	The framework used to implement writing, testing and running MapReduce pipelines.	Big-Data Library	DOAP RDF (json)	Java and Scala

Программное обеспечение Apache

DataFu	Consist of two library- pig and hourglass. This works for data mining and statistics	Big-Data Incubating	DOAP RDF (JSON)	Java
Drill	Query language as distributed SQL MPP with Hadoop and NoSQL	Big- data	DOAP RDF (json)	Java
Edgent	It is programming model used for streaming process to execute analytics	Big-Data, Library, Mobile network client	DOAP RDF (JSON)	Java, JavaScript
Falcon	Platform for Data management and processing	Big-Data Incubating	DOAP RDF (json)	Java
Flink	Rapid and trustworthy for voluminous scale data processing	Big-Data	DOAP RDF (JSON)	Java and Scala

Программное обеспечение Apache

Flume	**Flume is trustworthy, distributed, efficient, aggregation to store data in a centralized manner**	**Big-Data**	**DOAP RDF (JSON)**	**Java**
Giraph	Giraph is developed to high scalability and iterative graph processing	Big-Data	DOAP RDF (JSON)	Java
Hama	Hama consist of BSP computing engine	Big-data	DOAP RDF (json)	Java
Helix	Framework uses clustering analysis for data partition and replication data resources	Big-Data	DOAP RDF (JSON)	Java
Ignite	Ignite is In-Memory Data providing processing, querying components	Big-Data, SQL, Cloud, OSGi IoT	DOAP RDF (JSON)	Java, C#, C++, SQL, JDBC, and ODBC

Программное обеспечение Apache

Kafka	**Open source programming provides distributed fault tolerance**	**Big-Data**	**DOAP RDF (JSON)**	**Scala, Java**
Knox	API gateway to Hadoop service	Big-Data, Hadoop	DOAP RDF (JSON)	Java
Lens	Provides Unified Analytics interface	Big-Data	DOAP RDF (JSON)	Java
MetaModal	Put forth uniform connector, query API to various Datastore Types	Database, Big-Data library	DOAP RDF (JSON)	Java
Oozie	Workflow scheduler to access Hadoop jobs	Big-Data	DOAP RDF (JSON)	Java, JavaScript