

GROUP TASK-2

Big Data Process Mapping

Introduction

Big Data refers to extremely large datasets that cannot be managed, processed, or analyzed using traditional data processing tools. The purpose of this report is to map the end-to-end process of handling big data, from collection to analysis, to ensure proper management, usability, and insights generation.

Process mapping helps in visualizing workflows, identifying bottlenecks, improving efficiency, and maintaining data quality. It is essential for organizations to understand the flow of data to make informed decisions and optimize big data operations.

1. Data Sources

Google Maps collects data from diverse and distributed sources, making it a classic example of big data characterized by high volume, velocity, and variety.

Primary data sources include:

- **Smartphones and user devices:** GPS location, movement speed, navigation queries, and search patterns.
- **Satellite imagery:** High-resolution geographic images used for terrain and map generation.

2. Data Ingestion

The ingestion layer is responsible for collecting and transmitting raw data into Google's infrastructure.

Key ingestion mechanisms:

- Real-time streaming from mobile devices.
- API-based integration for third-party data.

3. Data Storage

Due to massive data scale, Google Maps relies on distributed and highly scalable storage systems.

Storage technologies include:

- Distributed file systems like Google File System (GFS) and Colossus.
- NoSQL databases such as Bigtable.

4. Data Processing :The processing layer transforms raw data into usable intelligence using distributed computing and AI.

Batch Processing:

- Map building and updates.
- Satellite image stitching.

Real-Time Processing:

- Live traffic congestion analysis.
- Incident detection and rerouting.

Key technologies and methods:

- Distributed computing models similar to MapReduce.
- Graph algorithms like Dijkstra and A* for shortest path routing.

Technologies Used

Data Collection & Ingestion: Apache Kafka, Apache Flume, Sqoop

Data Storage: Hadoop HDFS, Amazon S3, NoSQL Databases (MongoDB, Cassandra)

Data Processing & Transformation: Apache Spark, Hadoop MapReduce

Data Analytics & Visualization: Tableau, Power BI, Apache Hive, Jupyter Notebook

Data Security & Governance: Apache Ranger, Kerberos, Data Governance Tools

Programming Languages: Python, Java, Scala Version Control & Collaboration: Git, GitHub

Google Maps continuously improves through a feedback-driven learning system.

Process Mapping

Big Data Process Mapping outlines the flow of data from collection to analysis, showing how raw data is transformed into actionable insights. The process typically involves the following stages:

1. **Data Collection:** Gathering data from multiple sources such as IoT devices, social media, web logs, sensors, and transactional systems.
2. **Data Ingestion:** Loading raw data into storage systems using batch or real-time ingestion tools like Apache Kafka, Flume, or Sqoop.
3. **Data Storage:** Storing structured, semi-structured, and unstructured data in distributed storage systems such as Hadoop HDFS, Amazon S3, or NoSQL databases.

4. **Data Processing & Cleaning:** Transforming and cleaning data to remove errors, duplicates, and inconsistencies. Using tools like Apache Spark or Hadoop MapReduce for large-scale data processing.
5. **Data Analysis:** Applying statistical techniques, machine learning models, or analytics algorithms to extract insights. Using tools like Apache Hive, Python, or Jupyter Notebooks.

Challenges Faced

While working on big data processes, several challenges were identified:

1. **High Volume of Data:** Managing extremely large datasets required efficient storage and processing techniques.
2. **Data Variety:** Handling structured, semi-structured, and unstructured data from multiple sources was complex.
3. **Data Velocity:** Processing real-time streaming data with minimal latency posed performance challenges.
4. **Data Quality Issues:** Ensuring accuracy, consistency, and completeness of data during ingestion and processing was difficult.
5. **Integration of Heterogeneous Sources:** Combining data from different platforms, formats, and systems required careful mapping

Conclusion

The Big Data Process Mapping report highlights the complete lifecycle of big data, from collection and storage to processing, analysis, and visualization. By mapping each stage, organizations can gain a clear understanding of workflows, identify bottlenecks, and optimize data handling for better efficiency and accuracy.

This process ensures that large and complex datasets are managed securely, consistently, and effectively, enabling actionable insights and data-driven decision-making.