

• Apache Pig:

- Platform for analyzing large datasets
- High-level language for describing data analysis programs and infrastructure for evaluating these programs
- Two parts of Pig:
 - Pig Latin (programming language)
 - Pig Runtime (environment)
- Pig supports many data types (int, float, double, long, chararray, bytearray)
- Pig advanced data types are:
 - Map
 - Tuple
 - Bag
- Pig Latin can easily generate a sequence of MapReduce jobs in an abstract way.

• Apache Drill:

- Open-source application that can perform analysis on massive datasets while working in a distributed context
- Its main strength is to "drill" into the variety of NoSQL databases and file systems with only a single query.

• Apache Sqoop:

- Bidirectional data transfer of large datasets between Hadoop and relational database servers
- Data migration tool based on a connector architecture to connect with different databases
- Used to import/export data from a wide variety of data stores (HDFS, Hive, HBase, Hcatalog, Accumulo,

Oracle, My Sql, PostgreSQL, SQL server, DB2)

• Apache Avro:

- Open source remote procedure call and serialization framework
- Enables the interchange of large datasets between programs.
- Uses JSON format for representing data definition and binary format to store data.
- Features of Avro:
 - Rich data structures
 - Compact, fast, binary data format
 - Robust support for dynamically changing data schema
 - A container file, to store persistent data
 - Simple integration with dynamic languages.

• Apache Mahout:

- Framework to facilitate the free implementation of distributed and scalable machine learning algorithms
- Ready-to-use framework for fast and effective analysis of large volumes of data using:
 - Classification
 - Clustering
 - Recommendation
 - Matrix and vector libraries
 - Distributed fitness function capabilities for evolutionary programming.

Data Management Component

• Zookeeper :

- Open-Source, centralized service for coordinating numerous distributed services (manage configuration, naming, distributed synchronization, delivering group services).
- Zookeeper has a Hierarchical name-space that allows a file to also be a directory.
- Working of Apache Zookeeper :
 - The clients are the nodes that make service request to the server, and the server is the node that fulfills those service requests.
 - When the services first begin, a chief administrator is chosen from the servers.
 - Every client establish a connection with one server.
 - Once a server receives a request to "write", it's sent to the leader, who will then poll the quorum to see if it agrees with the action.

• ZNode :

- The nodes in a Zookeeper tree are referred to as ZN.
- The Responsibility of a ZNode is to keep track of the statuses of the structure as well as the version numbers for any data duplicate.
- Types of ZNodes :
 - Persistent ZNodes :
 - Default ZNode.
 - Only by making a call to Zookeeper's delete feature can it be removed from the system.

- Ephemeral CNodes:

- No Chkd and get deleted along with the data stored when a session is ended.

• Oozie:

- Scalable, reliable and extensible workflow scheduler to manage Apache Hadoop jobs.
- The default web server used in Oozie is Apache Tomcat.
- The Oozie Server is a stateless web application that does not retain any information in memory relating to the user or the job.
- The client can interact with the server in 3 ways:
 - Command line tool
 - Java client API
 - HTTP REST API

• Ambari:

- Easy-to-use Hadoop management web UI backed by RESTful APIs.
- Provision, manage and monitor Apache Hadoop clusters.
- Features of Ambari:
 - Easy Step-by-step guide for installation and user-friendly configuration
 - Immediate insights into the health of Hadoop cluster
 - Emergency alerts / notifications when a node goes down
 - Central management
 - Interactive dashboard
 - Authentication, authorization & auditing
 - Flexible and adaptive

Apache Flume :

- Collects, aggregates and moves enormous amounts of data ^{log}
- Dependable and always available
- Reliable and tolerant big data handling system.

→ Advantages of Flume:

- Highly Scalable, very dependable and highly fault-tolerant service
- Consistent flow of data transfer.
- Interface with a wide variety of applications (Kafka, HDFS and Thrift).

Apache Spark:

- Alternative framework to Hadoop.
- Process data in a distributed manner and is replacing MapReduce in many cases.
- Apache Spark is known for its speed and ability to temporarily store its findings in memory space.
- Supports Java, Python, Scala, and R, and there is native assistance for SQL, real-time data (data streaming), machine learning (ML) and graph processing (GP).

Big Data (Chapter 5 - Part 2)

NoSQL Databases

Introduction

- NoSQL databases, referred to as "NoSQL" or "Not only SQL", are non-tabular databases and store data differently than relational tables.
- Data in a NoSQL database system can be stored in a wide variety of non-relational formats (Structured, unstructured, semi-structured, polymorphic data) across several databases.
- NoSQL provides high scalability and availability so it's frequently utilized in environments dealing with large amount of data.

Features of NoSQL

- Non-Relational:
 - They don't produce tables with records that have a flat format and fixed column count.
 - Data stored in NoSQL need not to be normalized.
 - NoSQL doesn't support features like self-contained aggregates, BLOBs, referential integrity, specific query languages, query planners.
- Schema Free:
 - It does not require any data structure explanation.
 - NoSQL offers multiple file types for exchanging data that are all related to the same subject area.
- Simple API:
 - Simple NoSQL APIs offer users storage and querying options for the data that has been provided, together with user interfaces that are basic and easy to understand and use.

→ These APIs make it possible to do low-level activities such as processing data and selecting items from a list.

→ Protocols are based on text and frequently used in conjunction with HTTP, REST and JSON.

Distributed:

- The administration of large numbers of NoSQL databases can be performed in a decentralized approach
- As a direct consequence of this, there is a lower level of coordination and a greater degree of resource dispersion.

Difference Between NoSQL and SQL

<u>Feature</u>	<u>SQL Databases</u>	<u>No SQL Databases</u>
<u>Concept</u>	Relational db used for structured data	Non-relational db used for unstructured data.
<u>Schema</u>	Prior defined schema and use SQL	Dynamic schemas for unstructured data
<u>Scalability</u>	Vertically scalable, Handle the load by "scale up" with a bigger server	Horizontally Scalable, "scale out" across commodity servers
<u>Structure</u>	Table-based with fixed rows and columns	Document, key-value, graph or wide column stores.

Feature	SQL Databases	No SQL Databases
Preferable we	Multi-row transactions	Unstructured data like documents or JSON
Multi-record ACID transactions	Supported	Most of them do not support multi-record ACID transactions. But some of them do (MongoDB)
Performance	Slow query processing	Faster than SQL db
Joins	Required to perform various query operations	Not required

Types of NoSQL Databases

- Key-value pair based
- Column-oriented databases
- Document-oriented databases
- Graph-based

Key-Value Pair Based Storage

- Data is kept as pairs of keys and values.
- It's designed to handle large amounts of data and heavy loads.
- Store data as a hash table in which each key is unique and the value can be a JSON object, BLOB, text, ...
- Like a relational database with only 2 columns:
 - Key or attribute name

→ Value.

- This kind of database is used to store different data types (collections, dictionaries, associative, arrays).

Column-Oriented Databases

- Based on Google BigTable concept.
- Organized as a set of columns.
- Just like Key-value Stores, column-oriented databases have some basic NoSQL structure while also preserving a lot of flexibility.
- In a single table column names and formatting can differ from row to row and each column is stored separately on disk.
- The information contained in each column is processed individually.

Row:

- Each row contains a unique row identifier to identify the column data.
- Every row can have a different number of columns from other rows.

Column:

- Each column is contained in its row.
- Each column consists of a name value pair along with a timestamp of its creation date.

Column family:

- Consist of multiple rows.

Document-Oriented Databases

- Used to store semi-structured data and descriptions of that data in document format.
- Stores data in JavaScript Object Notation (JSON), Binary JavaScript Object Notation (BSON) or XML.
- Documents can be nested and specific elements can be indexed for faster querying.

Graph-Based Databases

- Use graph structures to store data elements and to define the relationships between stored data elements.
- Useful for identifying patterns in unstructured and semi-structured data content.
- Because graph-oriented databases store the relationships between data elements, they can support richer representation of data relationships.

Base Model of NoSQL

- Although some of the NoSQL databases (Apache's CouchDB, IBM's Db2) follow traditional ACID compliance, most of them are designed not to strictly follow the ACID rules and focus on availability even in the case of frequent failures.

Parameter	Key-Value Pair Based DB	Column-Oriented DB	Document-Oriented DB	Graph-Based DB
Structure	Simple structure of a unique key and its associated value	Tables where a row key can link to a large number of columns	Document-like structure which can be encoded in JSON format	Data elements are stored as nodes in a graph and the relations are edges
Use Case	Session management, application logs, managing shopping carts, recording click stream data	Recommendation engines, catalogs, fraud detection, event logging	Content management, monitor mobile and web app, web analytics, e-commerce	Social media platforms, reservation systems, customer relationship management
Popular Databases	Aerospike, DynamoDB, Redis, Riak	Accumulo, Amazon Simple DB, Cassandra, HBase, Hypertable	Couchbase Server, CouchDB, MarkLogic, MongoDB	AllegroGraph, IBM Graph, Amazon Neptune, ArangoDB, Neo4j

• Apache Pig:

- Platform for analyzing large datasets
- High-level language for describing data analysis programs and infrastructure for evaluating these programs
- Two parts of Pig:
 - Pig Latin (programming language)
 - Pig Runtime (environment)
- Pig supports many data types (int, float, double, long, chararray, bytearray)
- Pig advanced data types are:
 - Map
 - Tuple
 - Bag
- Pig Latin can easily generate a sequence of MapReduce jobs in an abstract way.

• Apache Drill:

- Open-source application that can perform analysis on massive datasets while working in a distributed context
- Its main strength is to "drill" into the variety of NoSQL databases and file systems with only a single query.

• Apache Sqoop:

- Bidirectional data transfer of large datasets between Hadoop and relational database servers
- Data migration tool based on a connector architecture to connect with different databases
- Used to import/export data from a wide variety of data stores (HDFS, Hive, HBase, Hcatalog, Accumulo,

Oracle, My Sql, PostgreSQL, SQL server, DB2)

• Apache Avro:

- Open source remote procedure call and serialization framework
- Enables the interchange of large datasets between programs.
- Uses JSON format for representing data definition and binary format to store data.
- Features of Avro:
 - Rich data structures
 - Compact, fast, binary data format
 - Robust support for dynamically changing data schema
 - A container file, to store persistent data
 - Simple integration with dynamic languages.

• Apache Mahout:

- Framework to facilitate the free implementation of distributed and scalable machine learning algorithms
- Ready-to-use framework for fast and effective analysis of large volumes of data using:
 - Classification
 - Clustering
 - Recommendation
 - Matrix and vector libraries
 - Distributed fitness function capabilities for evolutionary programming.