Open Source Software Paradigms

Lecture - 06

Big Data Tools and Office Suites

Lecturer: Biniam Behailu
Addis Ababa Science and Technology University

Learning Objectives

By the end of this lecture, you will be able to:

- Define Big Data using the essential 5 V's framework.
- Explain the core roles of key tools in the data pipeline, like Hadoop, Spark, and Kafka.
- Differentiate between the two main processing paradigms: Batch vs. Streaming.
- Evaluate the key benefits and challenges of using open-source software.
- Identify the leading open-source office suite and the critical importance of the Open Document Format.

Contents

What is Big Data?

- Big Data Refers to Large, Complex Datasets typically involving High Velocity Data Collection And Storage.
- Big data analytics is the process of collecting, analyzing, and extracting insights from massive, complex datasets

What is Big Data?

- The 5 V's of Big Data:
  - ☞ **Volume**: The sheer scale of data (Terabytes, Petabytes, Exabytes).
  - ☞ **Velocity**: The speed at which data is generated and processed (real-time streams).
  - ☞ **Variety**: The different types of data (structured, semi-structured, unstructured).
  - ☞ **Veracity**: The quality and trustworthiness of the data.
  - ☞ **Value**: The ultimate goal extracting meaningful insights.

The Big Data Explosion: Scale & Complexity

Petabyte Scale

By 2025, organizations routinely handle petabytes of data originating from IoT devices, social media feeds, and complex cloud applications.

Beyond Traditional Limits

Legacy databases fail to manage the velocity and variety. Big data tools are essential for processing diverse data structured and unstructured at scale.

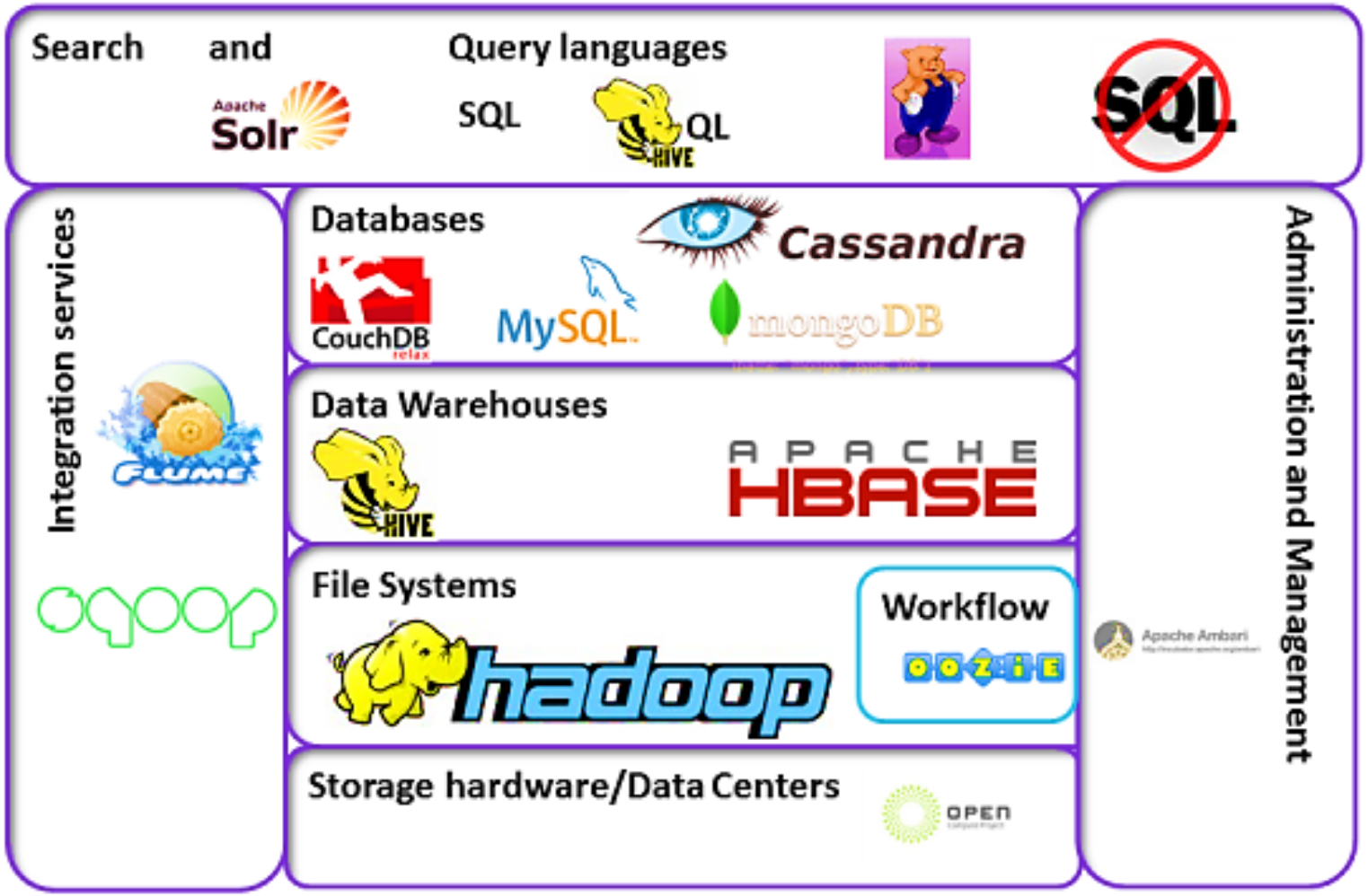
AI & Real-Time Necessity

Business operations now demand real-time analytics and AI-powered intelligence to maintain competitive relevance.

Why OSS for Big Data?

- Fosters rapid innovation and adoption.
- Avoids vendor lock-in.
- Drives industry standards (e.g., Hadoop, Spark).

Opensource Big Data Tools



OPENSOURCE BIG DATA TOOLS

***Image source**: R. Padaki, "Big Data Open Source Technology Landscape," Data Kulfi, 27-Mar-2013. [Online]. Available: https://datakulfi.wordpress.com/2013/03/27/big-data-open-source-technology-landscape/. [Accessed: 02-Oct-2025].*

The Foundational Model: Hadoop

- A framework that allows for the distributed processing of large data sets across clusters of computers.
- Includes Hadoop Distributed File System (HDFS) for storage and MapReduce for processing.
- Hadoop
  - ☞ Brings computation to the data.
  - ☞ Designed to run on commodity hardware.
  - ☞ Highly fault-tolerant.

Hadoop Ecosystem: Key Components

- **HDFS**: Reliable, scalable storage.
- **YARN** (Yet Another Resource Negotiator): The cluster resource manager.
- **MapReduce**: Batch processing engine.

Beyond Hadoop: Modern File Formats

- **Apache Parquet**: Columnar storage. Ideal for analytical queries.
- **Apache ORC**: Optimized Row Columnar format. Similar to Parquet.
- **Avro**: Row-based format, excellent for serialization and schema evolution.

Data Ingestion: Getting Data into the System

- **Apache Sqoop**: Designed for efficiently transferring bulk data between Hadoop and relational databases.
- **Apache Flume**: A service for efficiently collecting, aggregating, and moving large amounts of log data.
- **Apache Kafka**: A distributed event streaming platform for real-time data feeds. More than just ingestion.

Batch Processing Engines

- Processing data in large, discrete chunks.
- **Apache MapReduce**: The original. Powerful but slow due to disk I/O.
- **Apache Spark** (**Core**): In-memory processing. Up to 100x faster than MapReduce for certain tasks. Can handle both batch and streaming.

Stream Processing Engines

- Processing data in continuous, real-time streams.
- **Apache Storm**: One of the first real-time processing systems.
- **Apache Spark Streaming**: Uses micro-batches to provide streaming capabilities.
- **Apache Flink**: True streaming model with low latency and high throughput. Also handles batch as a special case of streaming.

Resource Management & Orchestration

- **YARN** (Hadoop): Manages and schedules resources (CPU, memory) across a Hadoop cluster.
- **Apache Mesos**: A general-purpose cluster manager that can handle diverse workloads.
- **Kubernetes** (CNCF): The de-facto standard for container orchestration. Now widely used for deploying and managing big data applications.

Coordination & Configuration

- **Apache Zookeeper**: A centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.

- The "**secret glue**" of the distributed systems world.

Workflow Scheduling

- **Apache Oozie**: A workflow scheduler system to manage Apache Hadoop jobs.
- **Apache Airflow**: A platform to programmatically author, schedule, and monitor workflows. More modern and feature-rich.

Cluster Management & Monitoring

- **Apache Ambari**: A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters.
- **Cloudera Manager** / **Hortonworks Data Platform** (**HDP**): Enterprise-grade cluster management tools (often include proprietary features).

Real-Time Search & Analytics

- **Elasticsearch**: A distributed, RESTful search and analytics engine built on Apache Lucene.
  - ☞ **Full-Text Search**: Powerful, fast, and relevance-based search across structured and unstructured text.
  - ☞ **Real-Time Analytics**: Instantly analyze and visualize your data.
  - ☞ **Log & Event Data Analysis**: Part of the popular "ELK Stack" (Elasticsearch, Logstash, Kibana).

Real-time OLAP Database

- **Apache Druid**: A high-performance, real-time, column-oriented, distributed data store designed for OLAP queries on event-driven data.
  - ☞ **Real-time Ingestion** & **Querying**: Ingest streaming data and immediately query it.
  - ☞ **OLAP Analytics**: Optimized for Online Analytical Processing with sub-second queries.
  - ☞ **Time-series Focus**: Native support for time-based data partitioning and queries.

Distributed SQL Query Engine

- **Presto**: A distributed, open-source SQL query engine designed for fast analytical queries against data of any size, from gigabytes to petabytes.
  - ☞ **ANSI-SQL Compliance:** Uses familiar SQL syntax, reducing learning curve.
  - ☞ **High Performance:** In-memory parallel processing for fast results.
  - ☞ **Flexibility:** Query data where it lives—no need for ETL into a single system.

Popular Big Data Tools Summary

Tool	Category	Use Case	Key Strength
HDFS	Storage	Distributed file storage	Fault-tolerant, scalable
Spark	Processing	Unified analytics (Batch/Streaming)	Speed, ease of use
Kafka	Ingestion	Real-time event streaming	High throughput, durable
Hive	Querying	SQL-on-Hadoop	Mature, SQL compatibility

Challenges of Open Source Big Data Tools

- **Complexity**: Steep learning curve for deployment and management.
- **Support**: Reliance on community forums; enterprise support may require paid subscriptions from vendors (e.g., Cloudera, Hortonworks).
- **Integration**: Making many different tools work together seamlessly can be difficult.
- **Security**: Requires careful configuration (Kerberos, Sentry, Ranger).

The Future of Big Data: Trends to Watch

- **Unification**: The rise of the Lakehouse architecture (Delta Lake, Apache Iceberg).
- **Real-time Everything**: Stream processing becoming the default.
- **MLOps**: Operationalizing machine learning models.
- **Serverless** & **Kubernetes**: Simplified deployment and management.
- **Data Mesh**: A decentralized, domain-oriented architectural paradigm.

Open Source Office Suites

What is an Office Suite?

- A collection of productivity software applications bundled together.
- Core Components:
  - ☞ Word Processor
  - ☞ Spreadsheet Application
  - ☞ Presentation Software

The Case for Open Source Office Suites

- **Cost**: Completely free to use and distribute.
- **Freedom**: No vendor lock-in or licensing fees.
- **Open Standards**: Strong support for Open Document Format (ODF), ensuring long-term accessibility.
- **Community-Driven**: Development is driven by user needs.

Popular Office Suites

1. LIBREOFFICE

2. OPENOFFICE

3. CALLIGRA SUITE

***Image source**: 1. Wikimedia Commons, "File:LibreOffice logo.svg," [Online]. Available: https://commons.wikimedia.org/wiki/File:LibreOffice_logo.svg. [Accessed: 02-Oct-2025].*

***Image source**: 2. Wikimedia Commons, "File:Apache OpenOffice logo and wordmark (2014).svg," [Online]. Available: https://commons.wikimedia.org/wiki/File:Apache_OpenOffice_logo_and_wordmark_%282014%29.svg. [Accessed: 02-Oct-2025].*

***Image source**: 3. Wikimedia Commons, "File:Calligra-logo.svg," [Online]. Available: https://commons.wikimedia.org/wiki/File:Calligra-logo.svg. [Accessed: 02-Oct-2025].*

LibreOffice: The Community Powerhouse

- **Origins**: Forked from OpenOffice.org in 2010.
- Key Features:
  - ☞ Active and vibrant community development.
  - ☞ Frequent releases and updates.
  - ☞ Extensive feature set.

Apache OpenOffice

- **Origins**: Descendant of IBM's and Sun Microsystems' StarOffice.
- Donated to the Apache Software Foundation in 2011.
- Key Features: Stable, mature codebase.

Calligra Suite

- Calligra Suite is a free and open-source office and graphics software suite developed by the KDE community.
- It was originally part of KOffice before being rebranded in 2010.
- User Base: Much smaller compared to LibreOffice (estimated <1% market share)
- Visibility: Less known outside of Linux/KDE enthusiast circles

File Format Focus: Open Document Format (ODF)

- An international standard (ISO/IEC 26300).
- XML-based, open, and royalty-free.
- Extensions: .odt (Text), .ods (Spreadsheet), .odp (Presentation).
- Ensures documents are accessible in the future, independent of any single vendor.

The Future of Office Suites

- **Cloud** & **Collaboration**: Moving towards web-based, real-time collaboration (e.g., Google Workspace model).
- **LibreOffice Online**: A cloud-based version of LibreOffice.
- **Deep Integration**: Tighter integration with other open-source tools and data sources.
- **AI Assistance**: Incorporation of AI for grammar checking, data analysis, and design suggestions.

Summary

- Addressed the challenges of Volume, Velocity, Variety, Veracity, and Value.
- Foundation & Evolution: Started with Hadoop (HDFS, MapReduce) and evolved into a rich ecosystem of specialized tools.
- The Right Tool for the Job:
  - ☞ **Ingestion**: Kafka for streams, Sqoop/Flume for bulk data.
  - ☞ **Processing**: Spark for fast batch & micro-batches, Flink for true streaming.
  - ☞ **Coordination**: Zookeeper as the "secret glue."
  - ☞ **Orchestration**: Airflow for managing complex workflows.
  - ☞ **Future** is **Unified** & **Real-Time**: Lakehouse architectures, Kubernetes, and Data Mesh are leading the way.

Summary

- Open Source Office Suites
  - ☞ **The Core Value Proposition**: Freedom from cost and vendor lock-in.
  - ☞ **LibreOffice is the Leader**: The most active, feature-rich, and community-driven suite.
  - ☞ **ODF is Key**: The Open Document Format ensures long-term accessibility and data sovereignty.
  - ☞ **Future** is **Collaborative** & **Intelligent**: Moving towards cloud-based collaboration and integrated AI assistance.

Brain Teaser

1. A large e-commerce company needs to process and analyze a week's worth of customer transaction data, which amounts to several petabytes, to generate its weekly sales report. Which data processing paradigm and engine are MOST suitable for this task?

A. Stream Processing with Apache Flink

B. Batch Processing with Apache Spark

C. Data Ingestion with Apache Kafka

D. Stream Processing with Apache Storm

Brain Teaser

1. A large e-commerce company needs to process and analyze a week's worth of customer transaction data, which amounts to several petabytes, to generate its weekly sales report. Which data processing paradigm and engine are MOST suitable for this task?

A. Stream Processing with Apache Flink

B. Batch Processing with Apache Spark

C. Data Ingestion with Apache Kafka

D. Stream Processing with Apache Storm

Brain Teaser

2. What is the primary strategic advantage of using the Open Document Format (ODF) in office suites like LibreOffice?

A. It allows for more advanced formatting than proprietary formats.

B. It is managed by a single vendor for consistent development.

C. It ensures long-term document accessibility and prevents vendor lock-in.

D. It is automatically compatible with all versions of Microsoft Word..

Brain Teaser

2. What is the primary strategic advantage of using the Open Document Format (ODF) in office suites like LibreOffice?

A. It allows for more advanced formatting than proprietary formats.

B. It is managed by a single vendor for consistent development.

C. It ensures long-term document accessibility and prevents vendor lock-in.

D. It is automatically compatible with all versions of Microsoft Word..

Thank you!

"Open source software is a testament to the power of collaboration; it transforms ideas into innovations, empowering individuals and communities to build a better future together."

Lecturer: Biniam Behailu
Addis Ababa Science and Technology University