The Hadoop Ecosystem

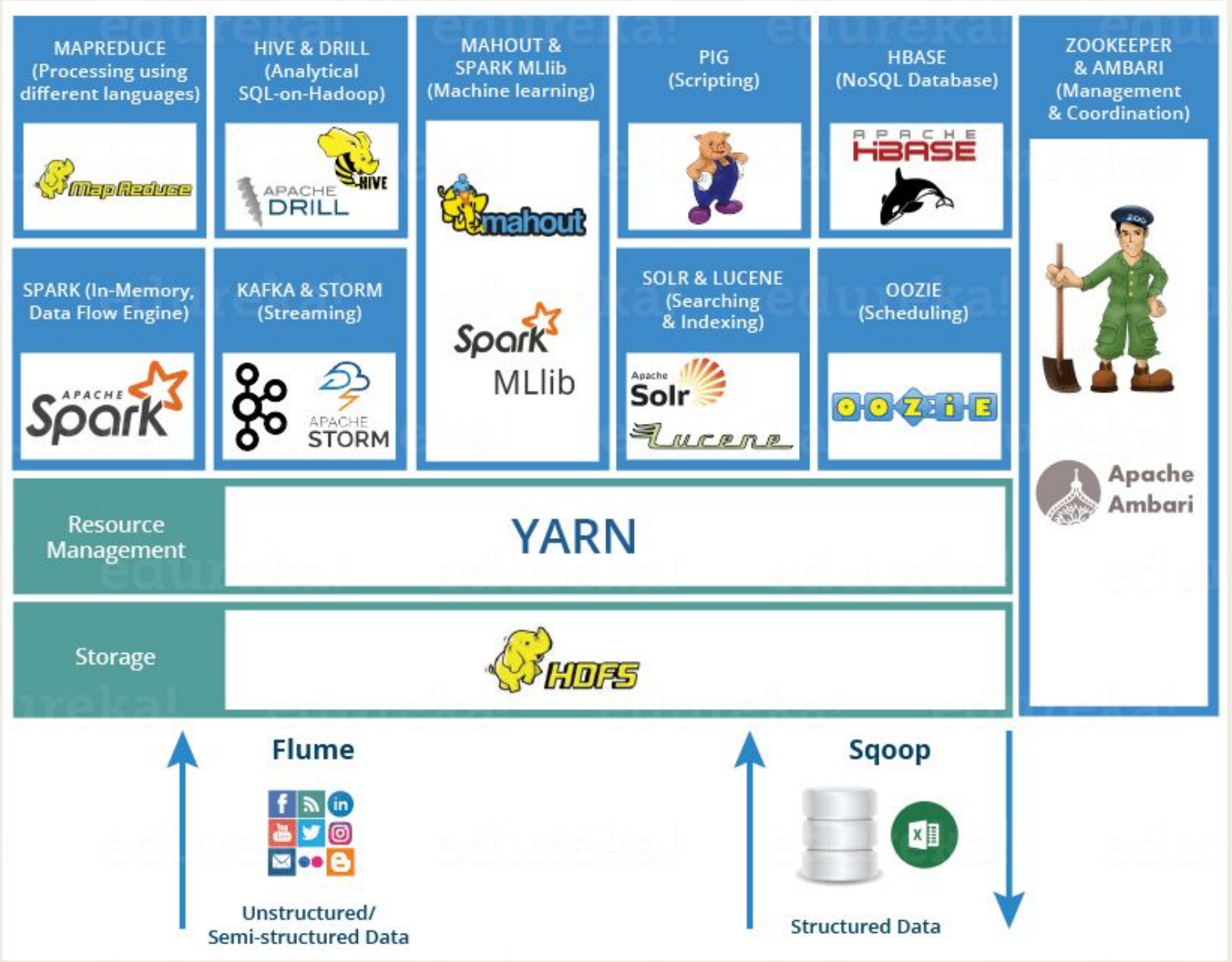
EECS 4415

Big Data Systems

Tilemachos Pechlivanoglou

tipech@eecs.yorku.ca

A lot of tools designed to work with Hadoop



HDFS, MapReduce

- Hadoop Distributed File System
  - Core Hadoop component
  - Distributed storage and I/O for Hadoop

- MapReduce
  - Core Hadoop component
  - Software framework for data processing

YARN

- Yet Another Resource Negotiator
  - Resource allocation and scheduling
  - Core Hadoop component

- Components: **ResourceManager**, **NodeManager**

  - **ResourceManager**:
    - receives processing requests
    - passes the parts of requests to corresponding NodeManagers
    - Has **Schedulers** that allocate resources, time based on application requirements
    - Has **ApplicationsManager** that monitors running jobs

  - **NodeManager**:
    - Handles requests at every DataNode

Apache Pig

- **SQL-like** command structure in Hadoop
  - Much more condensed (10 pig latin lines ≈ 200 Map-Reduce lines)
  - Allows actions like grouping, filtering etc.
  - Developed by Yahoo

- **Pig Runtime** and **Pig Latin** language
  - Analogy to Java: Pig Runtime -> JVM, Pig Latin -> Java
  - **Compiler** internally converts pig latin to MapReduce

Apache HIVE

- **SQL** queries in Hadoop:
  - Uses Hive Query Language(HQL), very similar to SQL
  - Highly scalable, both batch and real-time processing support
  - Supports all SQL types, most commands etc.

- **JDBC/ODBC driver** and **Hive Command Line** :
  - **Java Database Connectivity (JDBC), Object Database Connectivity (ODBC)**
    - Used to establish connection with data storage
  - Developed by Facebook

Apache Mahout

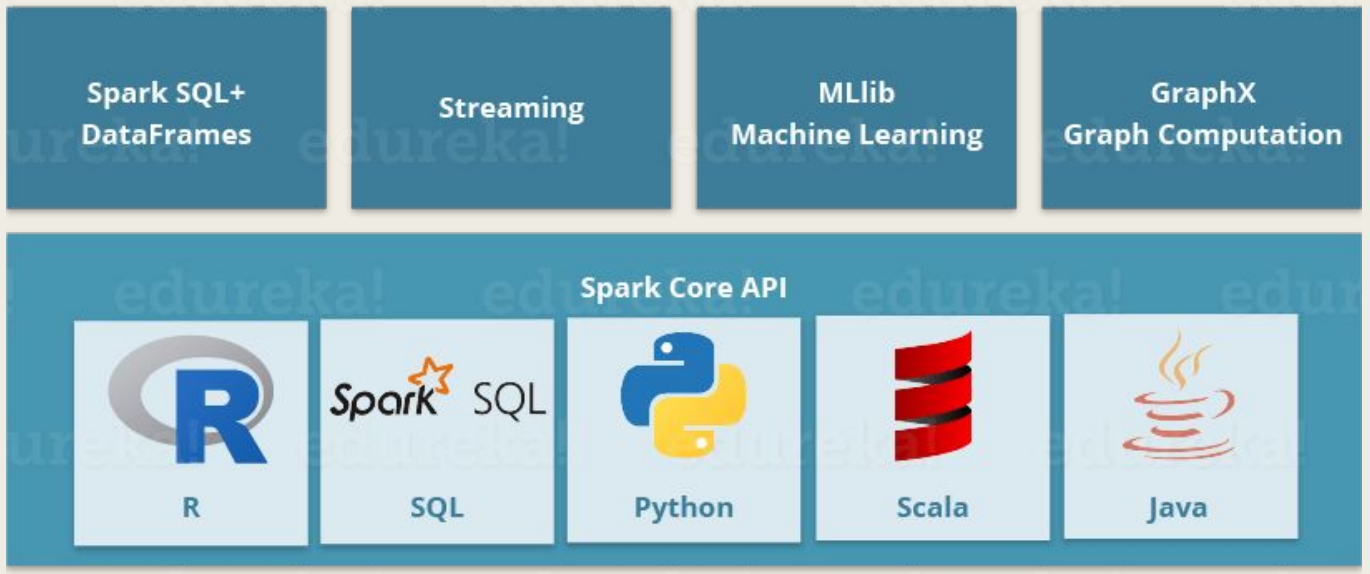
- **Machine Learning** in **Hadoop**
  - Provides built-in algorithms for machine learning problems
  - Executed through a command line

- Supported algorithms:
  - **Collaborative filtering**: mining patterns/behaviors, makes predictions and recommendations
    - Amazon product recommendation
  - **Clustering**: finding groups of similar data
    - recommending groups in social media
  - **Classification**: classifying and categorizing data into various sub-departments
    - identifying objects in image recognition

Apache Spark

- **Framework** for **real time** data analytics
  - Executes **in-memory** computations, **high-speed** data processing (100x faster than MapReduce)
  - Written in Scala, but supports many languages

- Contains high-level libraries, processing based on **DataFrames**



Apache HBASE

- **Non-relational** distributed database (**No-SQL**)
  - All types of data, absolutely everything is supported
  - Provides fault tolerance and fast retrieval of data
  - Open source, based on Google's BigTable

- Runs on top of Hadoop, provides BigTable - like capabilities
  - Written in Java

Apache Zookeeper, Oozie

- **Zookeeper**: Hadoop **job coordination**
  - Coordination between different **distributed** Hadoop **jobs/services**
  - Things like addresses, start-up/shutdown, configurations
  - Used in Rackspace, Yahoo, eBay

- **Oozie**: Hadoop clock/alarm
  - **Oozie Workflow**: sequential acts to be performed
  - **Oozie Coordinator**: triggers job execution when data is available

Apache Flume, Sqoop

- **Flume: Unstructured** data **ingestion**
  - Handles the entry of data in the system
  - **Collects, aggregates** and **moves** large amounts of data
  - Handles **real-time input streams**

- **Sqoop**: **Import/export structured** data
  - Also handles data ingestion
  - Moves data from **RDBMS** or **Enterprise** data warehouses to **HDFS** or vice versa

Apache Solr & Lucene

- **Searching** and **indexing**
  - Used for different data search tasks
  - Solr is the application, Lucene is the engine/kernel

Apache Ambari

- **Managing** the **whole ecosystem**

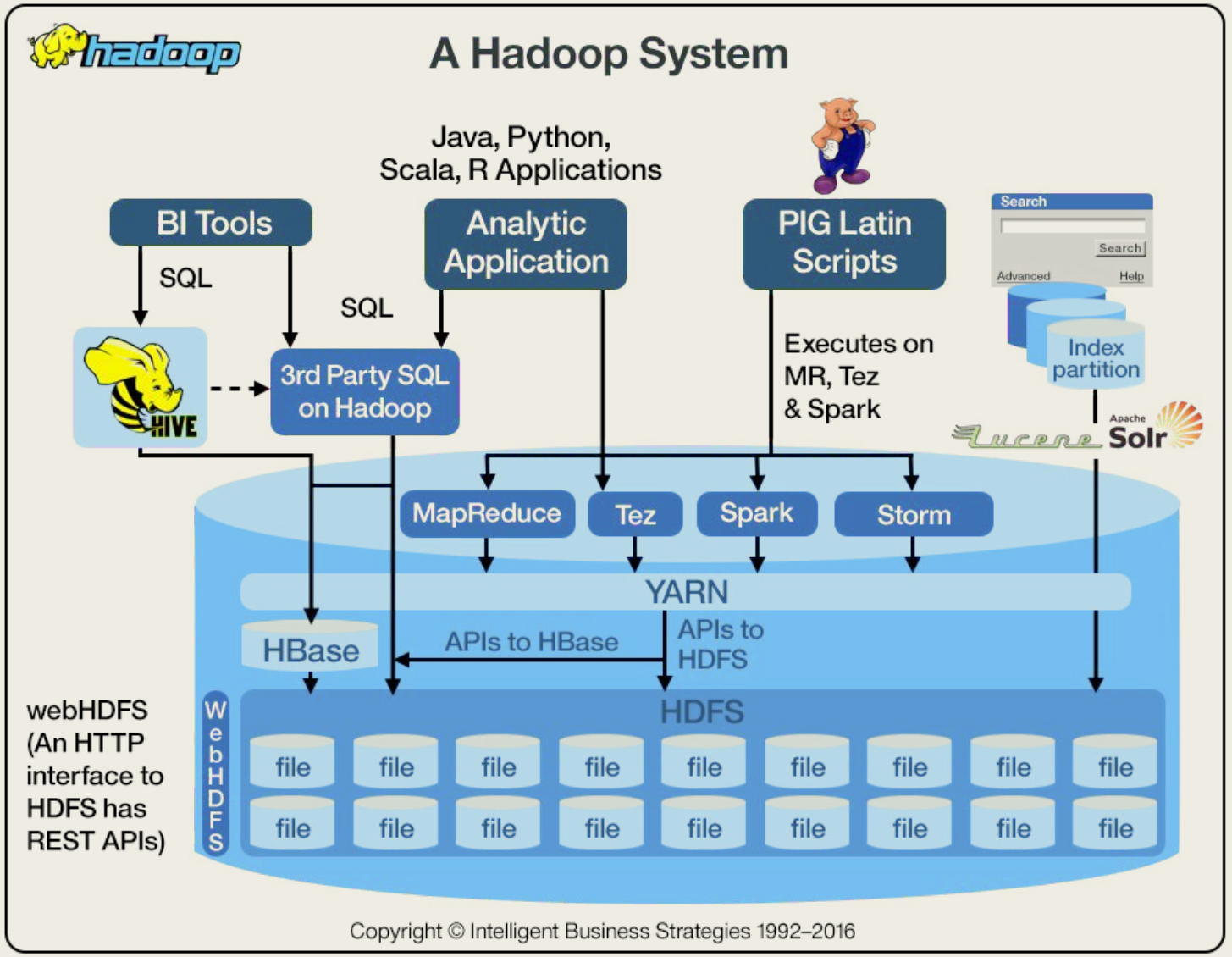
Apache Ambari

- **Hadoop cluster provisioning**
  - Step by step process for installing hadoop on many hosts
  - Handles Hadoop cluster configurations
- **Hadoop cluster management**
  - Provides central management service for starting, stopping and re-configuring Hadoop services
- **Hadoop cluster monitoring**
  - Dashboard for monitoring cluster health and status
  - Amber Alert framework for notifying if something is wrong

Honorable mentions

- **Avro**: data serialization (~JSON)
- **Cassandra**: reliable NoSQL distributed database
- **Cloudera**: Hadoop environment management, commercial vendor
- **Chukwa**: data collection system
- **Impala**: analytic database
- **Kafka**: Hadoop messaging
- **Tajo**: robust big data relational and distributed data warehouse
- **Tez**: generalized data-flow programming framework

An example Hadoop system



Thank you!

Based on:

https://www.edureka.co/blog/hadoop-ecosystem
http://www.bmc.com/guides/hadoop-ecosystem.html