SAKARYA ÜNİVERSİTESİ

BSM 461

INTRODUCTION TO BIG DATA

Lecture 9

Kevser Ovaz Akpınar, PhD

kovaz.sakarya.edu.tr
kovaz@sakarya.edu

Agenda

- Hadoop

Hadoop

- Open source software framework used for distributed storage and processing of big datasets
- Can be set up over a cluster of computers built from normal commodity hardware (difference of Hadoop from others)
- Many vendors offer their implementation of a Hadoop stack (e.g. Amazon, Cloudera, Dell, Oracle, IBM, Microsoft)

History of Hadoop

- 2 key building blocks:

-Google File System: A file system that could be easily distributed accross commodity hardware, while providing fault tolerance

-Google MapReduce: A Programming paradigm to write programs that can be automatically parallelized and executed accross a cluster of computers

-Doug Cutting developed Nutch web crawler!! Later named to Hadoop

-In 2008, Yahoo! Open sourced Hadoop as «Apache Hadoop»

Hadoop Stack

4 modules

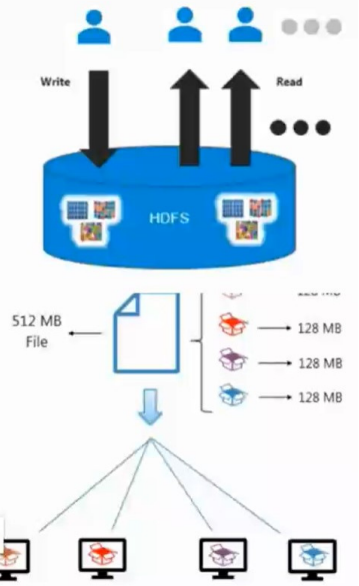
- Hadoop Common: Set of shared programming libraries used by the other modules
- HDFS (storage)
  - A Java-based file system to store data across multiple machines
  - Allows to dump any kind of data across the cluster
- MapReduce framework (processing)
  - A programming model to process large sets of data in parallel
  - Allows parallel processing of the data stored in HDFS
- YARN (Yet Another Resource Negotiator)

  Handles the management and scheduling or resource requests

  3 services on YARN->ResourceManager, JobHistoryServer, NodeManager

HDFS

- Storage unit of Hadoop
- It is a Distributed File System
- Divide files (input data) into smaller clunks and stores it across the cluster
- Scalable as per requirement
- Allows any kind of data, be it structured, semi-strucuted or unstructured
- Follows WORM (write one read many)
- No schema validation is done while dumping data



6 HDFS Services

- NameNode-> master for storage
- DataNode-> slave for storage
- Secondary NN ->backup for NN
- ResourceManager->Master for YARN
- NodeManager->Slave for YARN
- JobHistoryServer->Status archival for MR job

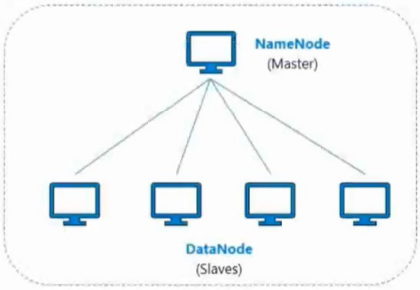
- Data blocks processed by NN and DN are typically 64MB

Java processes: Sudo /usr/java/latest/bin/jps

HDFS

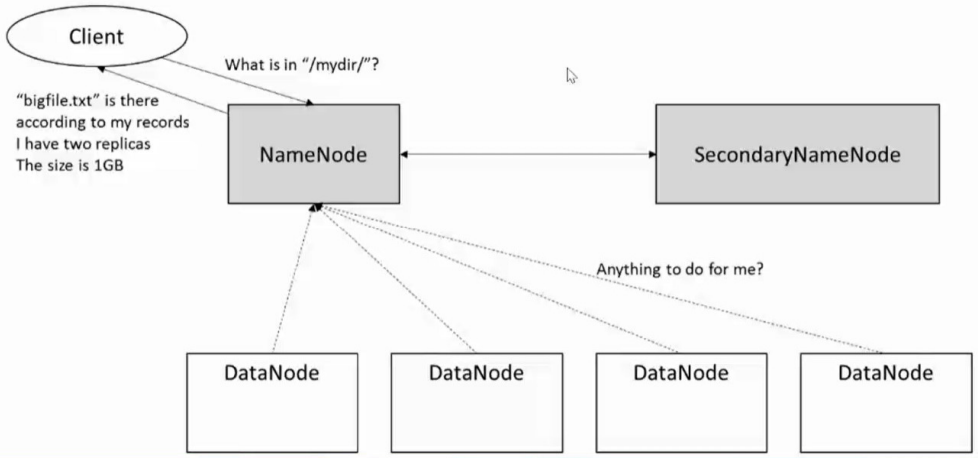
- HDFS is a distributed file system to store data across a cluster of commodity machines
- High emphasis on **fault-tolerance**
- It creates a level of abstraction over the resources, from where we can see the whole HDFS as a single unit
- 2 components (master–slave manner)
  - ***NameNode***
    - Main node that contains metadata about the data stored (which data blocks is stored,in which data node,where are the replications..)
    - Manages incoming file system operations (open,close,renaming..)
    - Maps data blocks (parts of files) to DataNodes
  - ***Datanodes*** are commodity hardware in the distributed environment
    - Handles file read and write requests
    - Create, delete and replicate data blocks amongst their disk drives
    - Continuously loop, asking the NameNode for instructions

Note: size of 1 data block is typically 64 MB

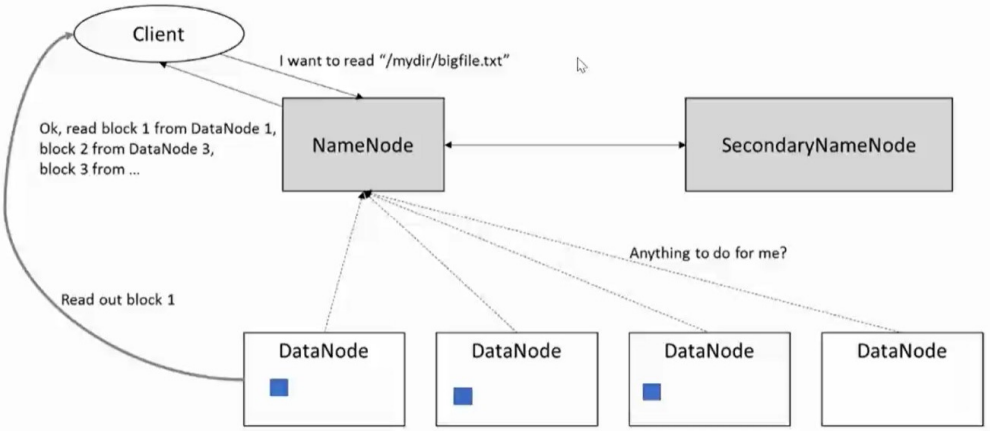


Hadoop Cluster

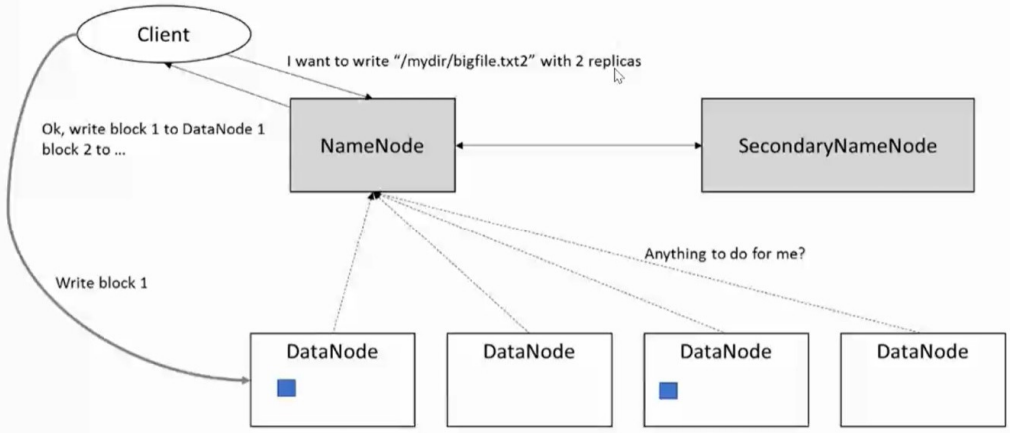
HDFS



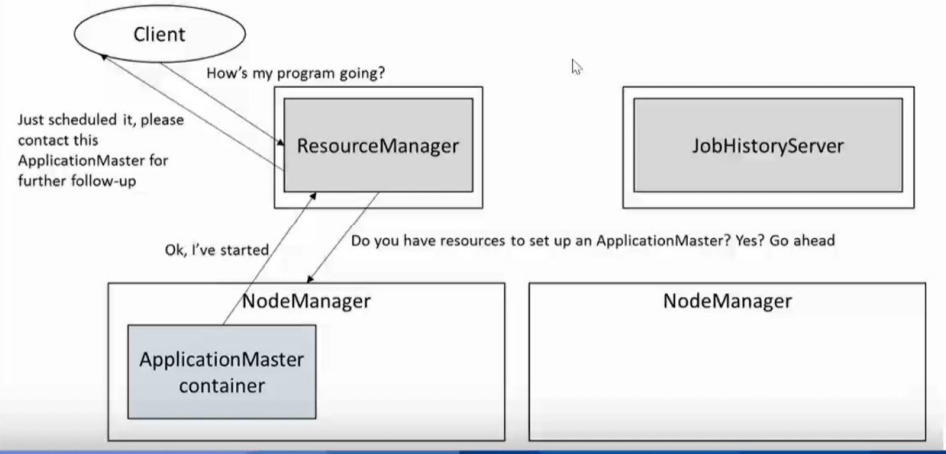
HDFS



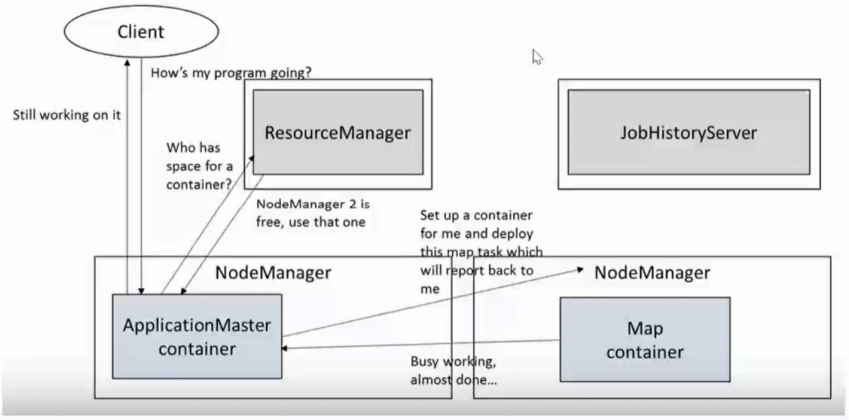
HDFS



YARN Services



YARN Services



HDFS Services

MapReduceAppX.java

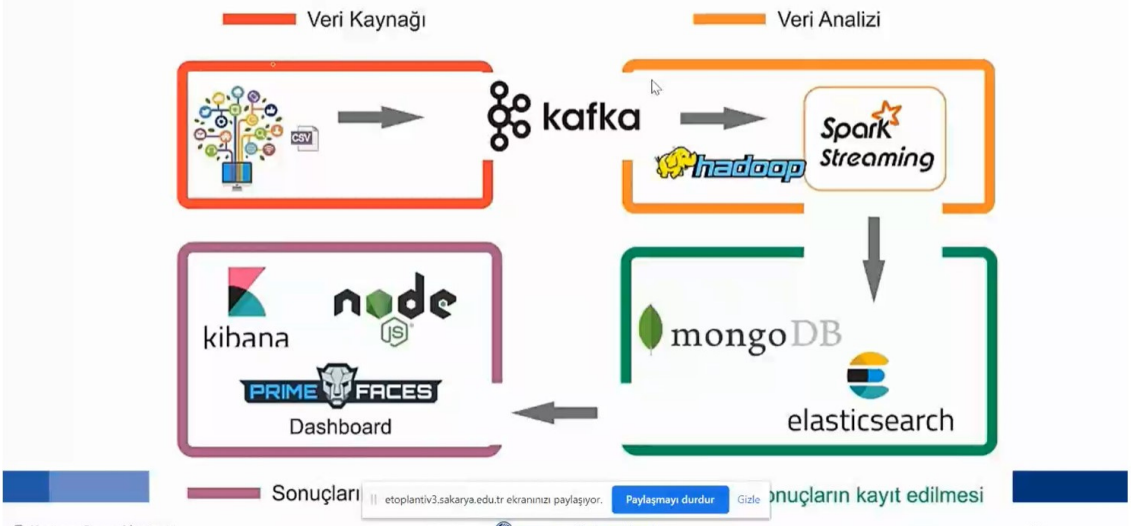
↓

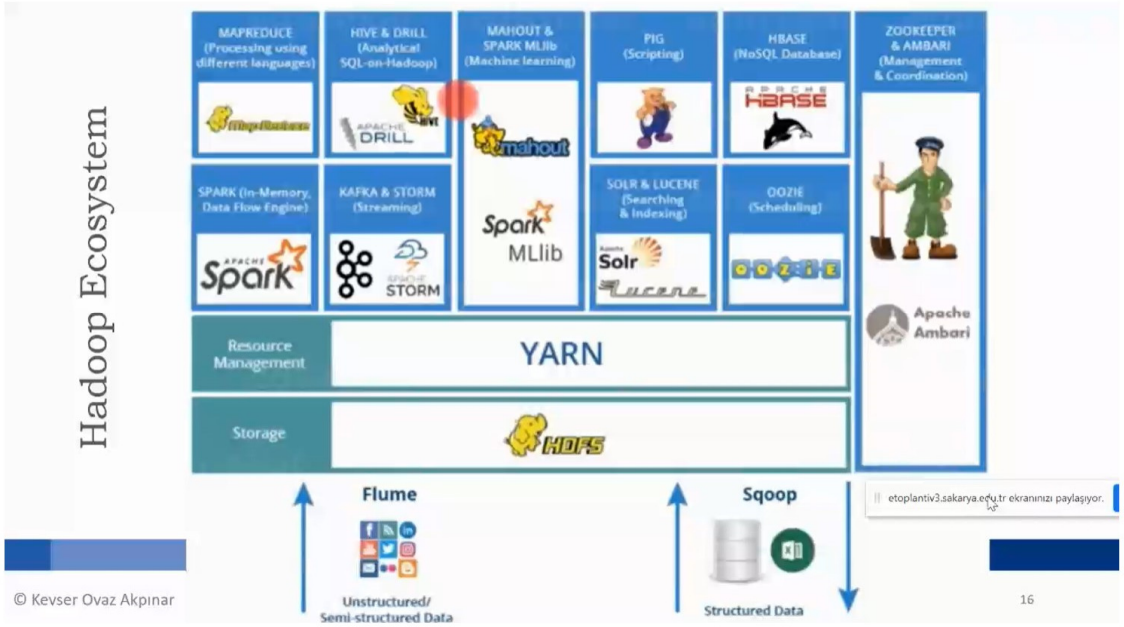
ResourceManager (port:8088)

↓

JobHistoryServer (port:19888)

Project Management Architecture





- Data Access: Pig, Hive
- Data Storage: HBase, Cassandra
- Interecation, Visualization, Execution,
- Development : HCatalog, Lucene, Hama, Cru
- Data Serialization: Avro, Thrift
- Data Intelligence: Drill, Mohout
- Data Integration: Sqoop, Flume, Chuwka
- Management: Ambari(Portal)
- Monitoring: Zookeeper
- Orchestration: Oozie
- **Hadoop on Browser:** HUE