HADOOP

Engineering Guild

Jakub Horcicka

02. 12. 2020

Agenda

Part I

- Hadoop, HDFS, MapReduce.
- YARN, Mesos, Spark.
- Flume, Sqoop, Hbase, Hive, Pig, Oozie, Hue.

Part II

- Data Scientist Workbench.
- MapReduce in java.
- Spark, Hbase, Oozie, Hue, DataCleaner on Spark (?).

Hadoop

Characteristics

- Open-source framework.
- Distributed parallel file system.
- Big data processing.
- Clusters.

History

- 2003 Google File System (paper).
- 2004 MapReduce (concept).
- 2006 Hadoop.
- 2008 Yahoo, Cloudera (distributor).
- Apache, Facebook, LinkedIn, eBay, IBM, . . .

Hadoop

Modules

- Hadoop Common.
  - Base for other modules.
- Hadoop Distributed File System (HDFS).
  - High throughput access to data.
- Hadoop YARN.
  - Job scheduling and resource management.
- Hadoop MapReduce.
  - Parallel processing of large data sets.

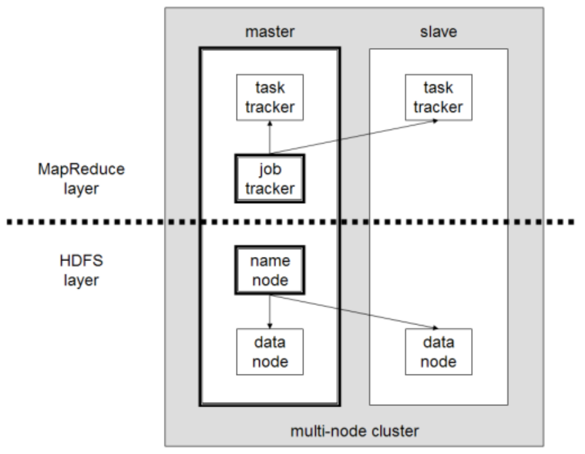
Hadoop

Architecture

- Node.
  - Single Name node.
    - Metadata: Directory tree, files list, no real data.
    - Rack awareness: Strategy to select nearest data node.
  - Multiple data nodes.
- Tracker.
  - Single Job tracker.
  - Multiple task trackers.
- Cluster.
  - Single Master & multiple slaves.

Hadoop

Architecture



Terms

Hadoop Distributed File System (HDFS)

- Designed to run on low-cost hardware.
- Computation distribution is more effective than data distribution.
- Fault tolerant (data replication).
- Batch processing / streaming access.
- Large data sets.
- NameNode (metadata) for each cluster.
- Data nodes → racks → cluster.
- Hundreds of nodes in a single cluster.

Terms

HDFS

- DEMO.

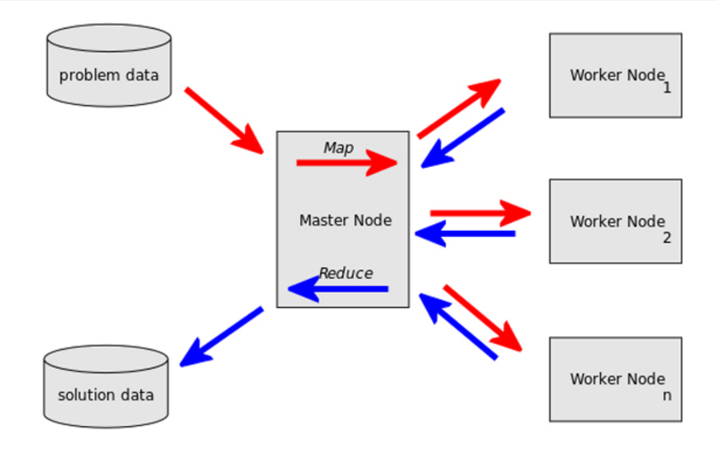
Terms

MapReduce I

- Programming model for big data (TBs) parallel processing.
- Input data are processed by map-tasks.
- Results are combined by reduce-tasks.
- Scheduling, monitoring.
- Single JobTracker, multiple TaskTrackers.
- Client submits a job (jar) and configuration to JobTracker.
- (MapR is a company that provides a Hadoop distribution).

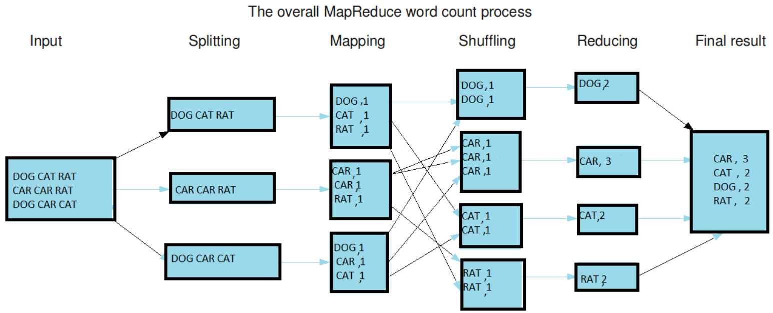
Terms

MapReduce II



Terms

MapReduce III



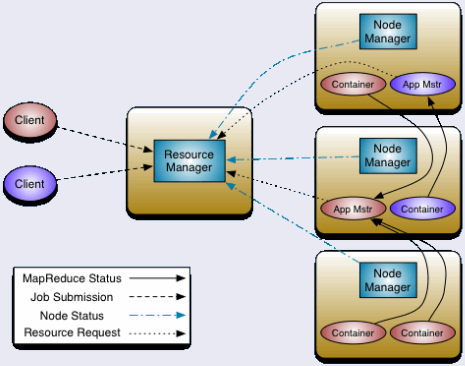
Terms

YARN (Yet Another Resource Negotiator)

- Operating system for big data apps.
- Multi tenancy (parallel jobs).
- Resource management.
- Hadoop2 (MapReduce2).
- Decouples resource management from scheduling.

Terms

YARN Scheduler and Application Managers



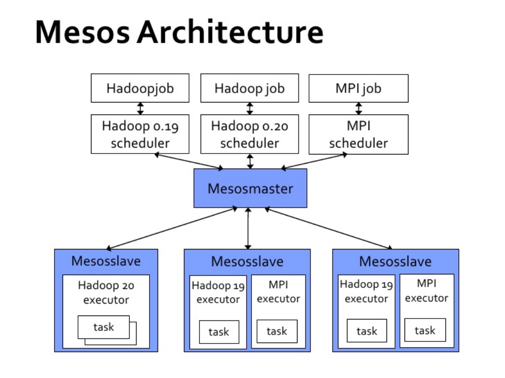
Terms

Mesos

- Cross-platform kernel for distributed systems (clusters).
- Provides API.
- Resource management (CPU, RAM, . . .).
- Scheduling.

Terms

Mesos architecture



Mesos Architecture

Terms

YARN vs. Mesos

- They can run parallelly next to each other.
- Main difference is a scheduler.

Hadoop YARN

- Improvement (next version) of MapReduce API.
- Best suited for Hadoop jobs.
- Job request → resource manager → evaluation → assignment.
- Server decides.

Apache Mesos

- Global resource manager (entire data center).
- Job request → master → offers → acceptance.
- Better scaling capabilities.
- Client decides.

Terms

Spark

- Open source cluster computing framework.
- Data structure oriented API.
- RDD (Resilient Distributed Dataset).
- Transformations and Actions.
- Shared memory.
- Database-style querying.
- Machine-learning algorithms.
- Requires a cluster manager (Spark cluster, YARN, Mesos).
- Requires a distributed storage system (HDFS, MapR-FS, Cassandra, . . .).

Terms

Spark parts

- Spark Core.
  - Tasks, scheduling, I/O, lazy-evaluated RDDs.
  - Java, Python, Scala, R.
- Spark SQL.
  - DataFrames abstraction layer.
  - Structured and semi-structured data.
- Streaming.
  - Analytics on mini-batches.
  - Support: Kafka, Flume, Twitter, TCP/IP sockets, ...
- Mlib.
  - Statistics, sampling, data generation, classification, ...
  - Cluster analysis methods (k-means).
- Graphx.
  - Graph (edges, vertices) processing framework.
  - Based on RDD.

Terms

Spark with DataCleaner

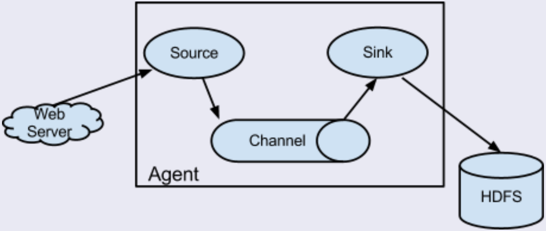
- DEMO.

Terms

Flume

- Log data collecting tool.
- Streaming.

Flume architecture



Terms

Flume

- DEMO.

Terms

Sqoop

- Data transfer between Hadoop and relational database.
- CLI tool.
- Import: Hive, HBase.
- Export: Hadoop → relational DB.

Terms

Sqoop

- DEMO.

Terms

HBase

- Hadoop database.
- Non-relational big-data store.
- Key-value.
- Real-time access.
- No sql-scripting.
- Java API.

Terms

Hive

- Data warehouse.
- Sumarization.
- Queries.
- Analysis.
- SQL-like interface (HiveQL).
- Command line tool.
- Java API.

Terms

Hive

- DEMO.

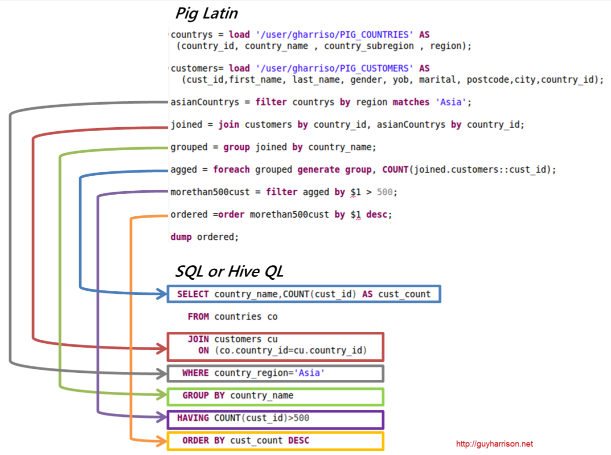
Terms

Pig

- Large datasets analyzing platform.
- Language Pig Latin.

Terms

Pig



Pig Latin

```
countrys = load '/user/gharriso/PIG_COUNTRIES' AS
 (country_id, country_name , country_subregion , region);

customers= load '/user/gharriso/PIG_CUSTOMERS' AS
  (cust_id,first_name, last_name, gender, yob, marital, postcode,city,country_id);

asianCountrys = filter countrys by region matches 'Asia';

joined = join customers by country_id, asianCountrys by country_id;

grouped = group joined by country_name;

agged = foreach grouped generate group, COUNT(joined.customers::cust_id);

morethan500cust = filter agged by $1 > 500;

ordered =order morethan500cust by $1 desc;

dump ordered;
```

SQL or Hive QL

```
SELECT country_name,COUNT(cust_id) AS cust_count

  FROM countries co

  JOIN customers cu
    ON (co.country_id=cu.country_id)
 WHERE country_region='Asia'
 GROUP BY country_name
HAVING COUNT(cust_id)>500
 ORDER BY cust_count DESC
```

http://guyharrison.net

Terms

Oozie

- Workflow scheduler system.
- Triggered by time, frequency or data availability.
- Jobs (DAG of actions).
- XML.

Terms

Hue

- Web interface.
- SQL editor for Hive.
- Searching.
- Spark and Hadoop notebooks.
- Job scheduling (Oozie).

Skillsoft courses

Part 1/4

- The Big Data Technology Wave
- Big Data Opportunities and Challenges
- Programming and Deploying Apache Spark Applications
- Apache Hadoop
- MapReduce Essentials
- Ecosystem for Hadoop
- Installation of Hadoop
- Data Repository with HDFS and HBase
- Data Repository with Flume
- Data Repository with Sqoop

Skillsoft courses

Part 2/4

- Data Refinery with YARN and MapReduce
- Data Factory with Hive
- Data Factory with Pig
- Data Factory with Oozie and Hue
- Data Flow for the Hadoop Ecosystem
- Designing Hadoop Clusters
- Hadoop in the Cloud
- Deploying Hadoop Clusters
- Hadoop Cluster Availability
- Securing Hadoop Clusters

Skillsoft courses

Part 3/4

- Operating Hadoop Clusters
- Stabilizing Hadoop Clusters
- Capacity Management for Hadoop Clusters
- Performance Tuning of Hadoop Clusters
- Cloudera Manager and Hadoop Clusters
- Big Data Corporate Leadership Perspective
- Big Data Engineering Perspectives
- Big Data - The Legal Perspective
- Big Data Marketing Perspective
- Big Data Strategic Planning

Skillsoft courses

Part 4/4

- Big Data Sales Perspective
- Spark Core
- Spark Streaming
- MLlib, GraphX, and R

References, links

- *https : //en.wikipedia.org/wiki/Apache*$_{H}$*adoop*
- *https : //en.wikipedia.org/wiki/Apache*$_{S}$*park*
- *http : //hadoop.apache.org/*
- *http : //spark.apache.org/*
- *http : //mesos.apache.org/*
- *http : //sqoop.apache.org/*
- *https : //flume.apache.org/*
- *https : //hbase.apache.org/*
- *https : //hive.apache.org/*
- *https : //pig.apache.org/*
- *http : //oozie.apache.org/*
- *http : //gethue.com/*
- *https : //courses.bigdatauniversity.com*

That's all, folks!

Thank you for your attention.