Hadoop YARN

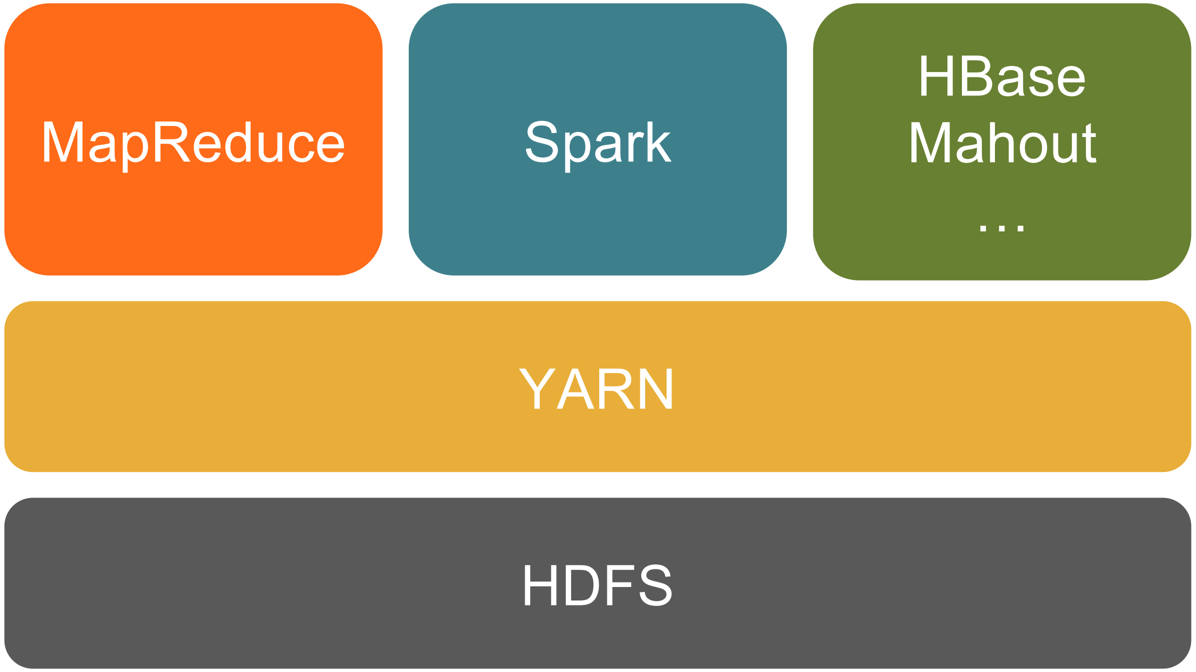
Amy Krause, Andreas Vroutsis

EPCC, The University of Edinburgh

Outline

- Hadoop ecosystem
- Hadoop YARN
- Hadoop use cases

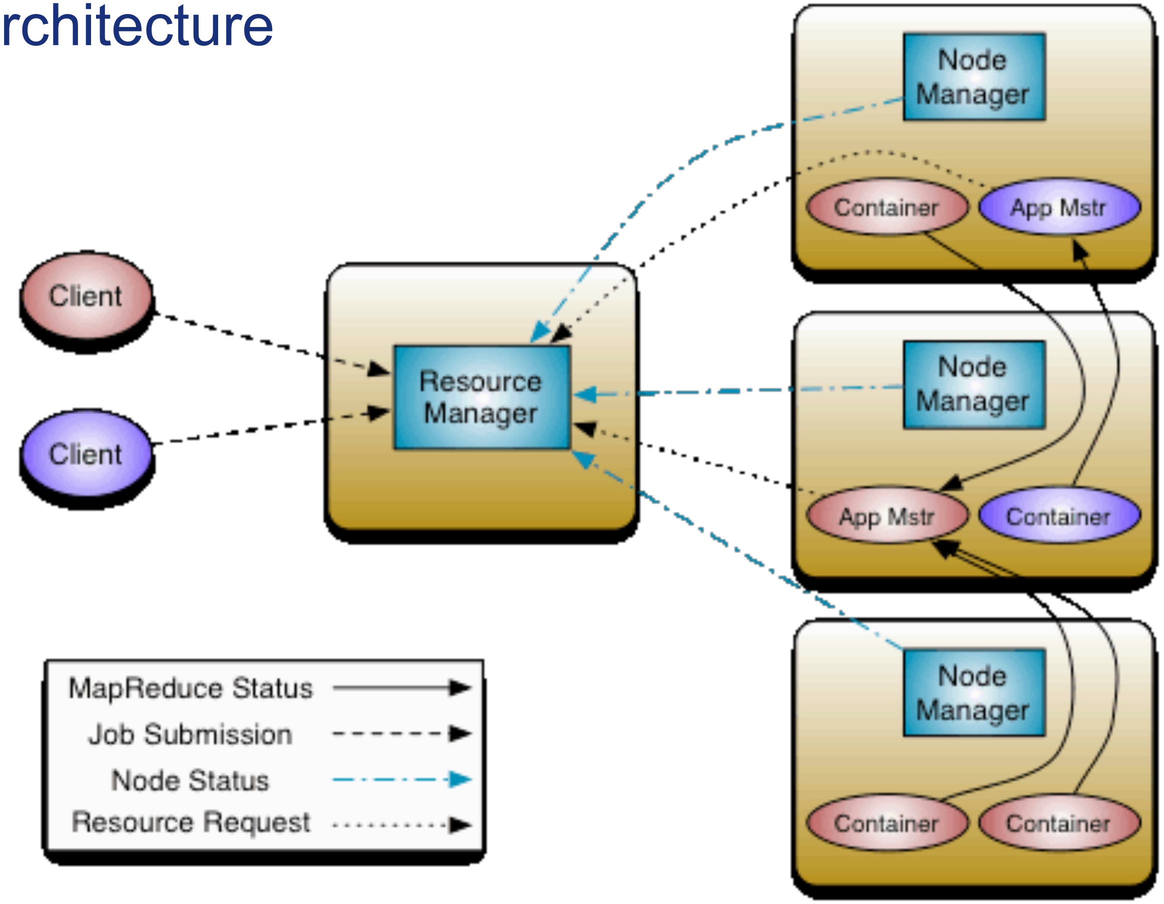
Hadoop system



Hadoop YARN overview

- YARN is a job scheduler
- It manages *resources* (nodes)
- It manages *applications* (a single job or a DAG of jobs)
- The **Application Manager** negotiates resources with the **Resource Manager** and works with the **Node Manager** to execute and monitor tasks

YARN Architecture



YARN Components

- **ResourceManager**
  - **Scheduler**: assigns cluster resources to queues and applications
  - **ApplicationsManager**: accepts job submissions and creates the ApplicationMaster
- **NodeManager**:
  - responsible for launching and managing containers
  - Manages the “health” of the node and reports to ResourceManager
  - **ApplicationMaster**
    - specifies the tasks to execute in a container
  - **Container**

Hadoop Applications

- Financial
- Retail
- IoT
  - Fast processing of volatile and large volumes of data
  - Reliable, robust and scalable for peak demand
  - Storage and analytics
- Data Analytics
  - Structured, semi-structured and unstructured data
- Data storage
  - Data lakes

Hadoop applications

- Hundreds of millions of Tweets are processed, stored, cached, served and analysed every day
- Hadoop clusters are running both compute and HDFS
- “We have multiple clusters storing over 500 PB divided in four groups (real time, processing, data warehouse and cold storage). Our biggest cluster is over 10k nodes. We run 150k applications and launch 130M containers per day.”
- Optimised data storage and workflow solutions within Hadoop

https://blog.twitter.com/engineering/en_us/topics/infrastructure/2017/the-infrastructure-behind-twitter-scale.html
https://github.com/twitter/elephant-bird

Hadoop applications

- “Expedia provisions Hadoop clusters using Amazon Elastic Map Reduce (Amazon EMR) to analyze and process streams of data coming from Expedia’s global network of websites, primarily clickstream, user interaction, and supply data"

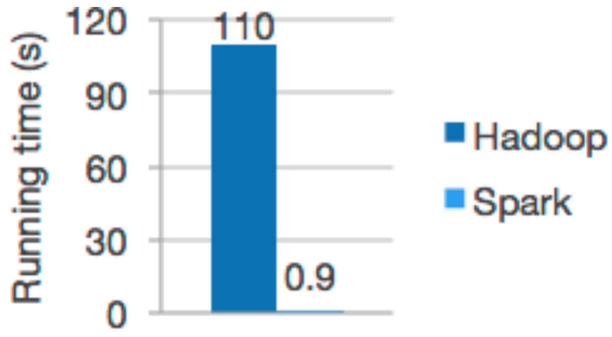
https://aws.amazon.com/solutions/case-studies/expedia/

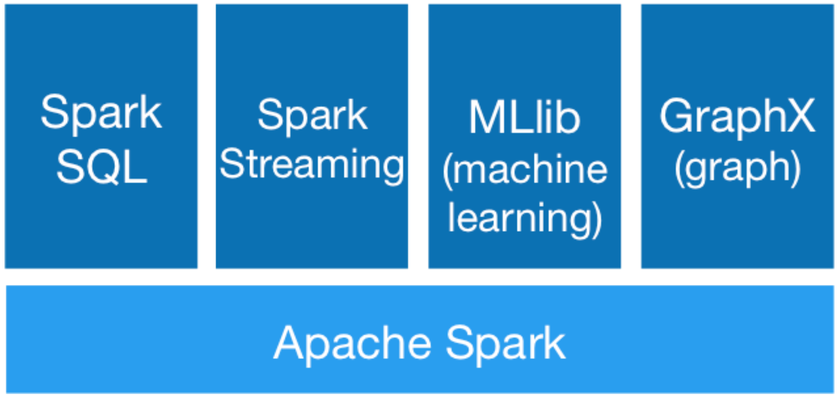
BRITISH AIRWAYS

… and many more!

A little more on Spark

- Explicitly supports caching data
  - Speeds up iterative algorithms
- Can use HDFS as the data source
- More that just map/reduce
  - Transformations:
    - map, filter, union, Cartesian, join, sample…
  - Actions:
    - reduce, collect, count, first, countBy, foreach…




Spark SQL

Spark Streaming

MLlib (machine learning)

GraphX (graph)

Apache Spark

THANK YOU FOR YOUR ATTENTION

www.prace-ri.eu