Big Data Architectures
For Machine Learning and Data Mining

Seminar Kick-Off Meeting

April 6th 2016

Web Technology and Information Systems Group

Michael Völske
michael.voelske@uni-weimar.de

What is Big Data

Different Points of View

> “Big data” is data that can’t be processed using standard databases because it is **too big, too fast-moving, or too complex** for traditional data processing tools.
>
> *AnnaLee Saxenian (Dean, UC Berkeley School of Information)*

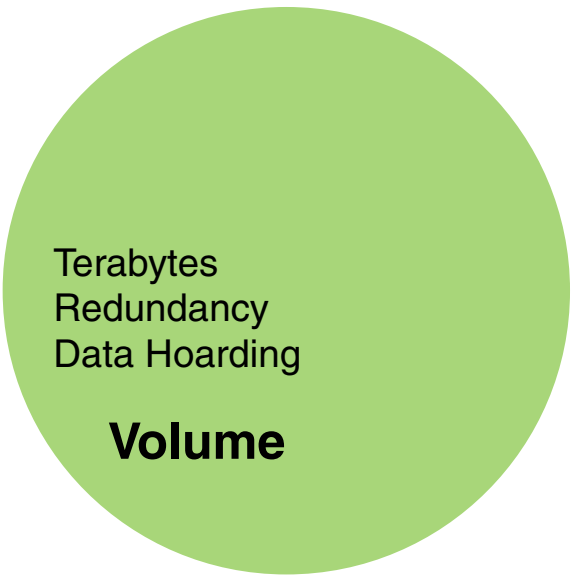
> Big data is when data grows to the point that the technology supporting the data has to change. It also encompasses a variety of topics relating to **how disparate data can be combined**, processed into insights, and/or reworked into smart products.
>
> *Anna Smith (Analytics Engineer, Rent the Runway)*

> In my view, big data is data that requires novel processing techniques to handle. Typically, **big data requires massive parallelism** in some fashion (storage and/or compute) to deal with volume and processing variety.
>
> *Brad Peters (Chief Product Officer, Birst)*

[http://datascience.berkeley.edu/what-is-big-data/]

What is Big Data

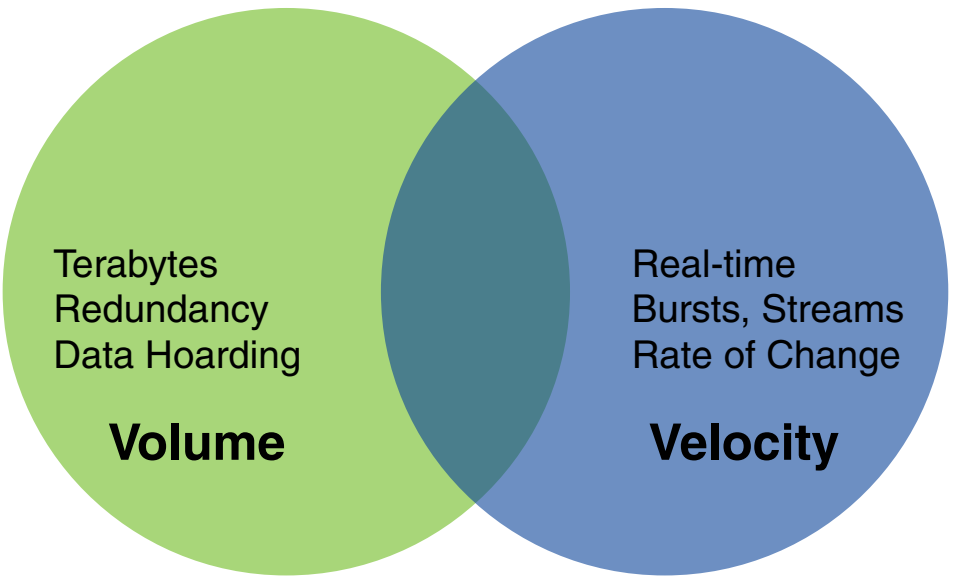
Gartner's "Three V's"



[http://www.gartner.com/it-glossary/big-data/]

What is Big Data

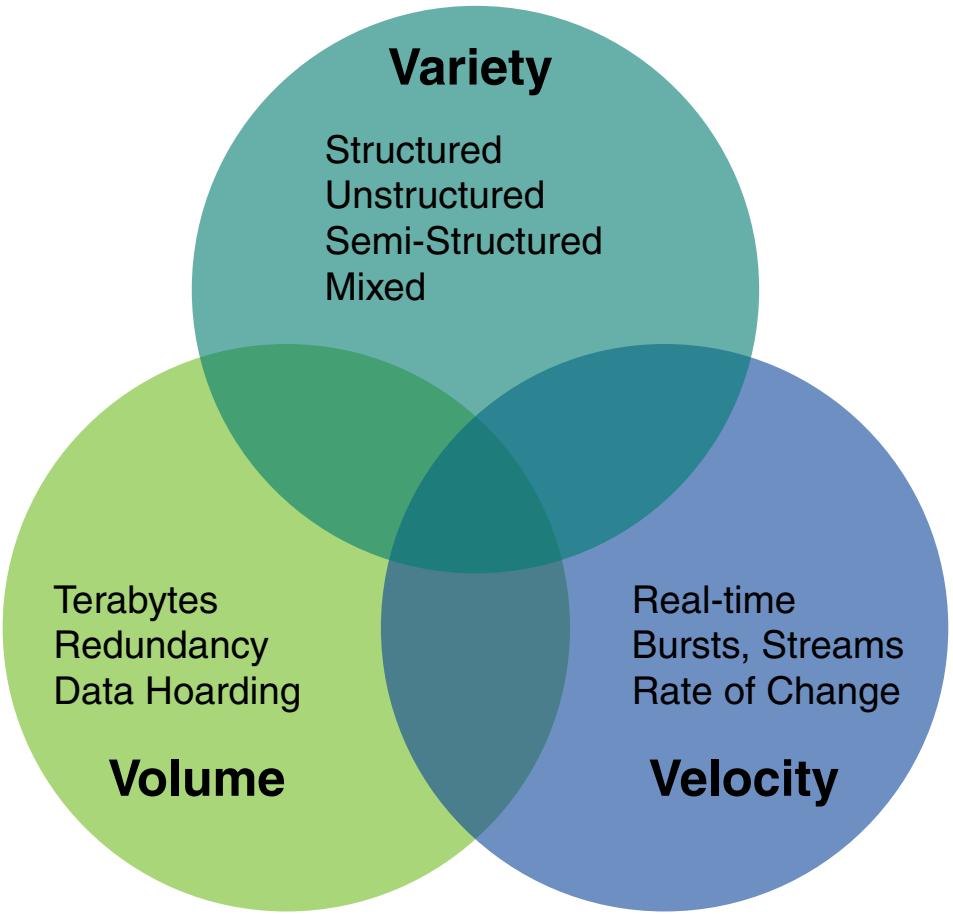
Gartner's "Three V's"



[http://www.gartner.com/it-glossary/big-data/]

What is Big Data

Gartner's "Three V's"



[http://www.gartner.com/it-glossary/big-data/]

The Big Data Architecture Stack

Data
Consumption
Layer

Data
Analytics
Layer

Data
Management
Layer

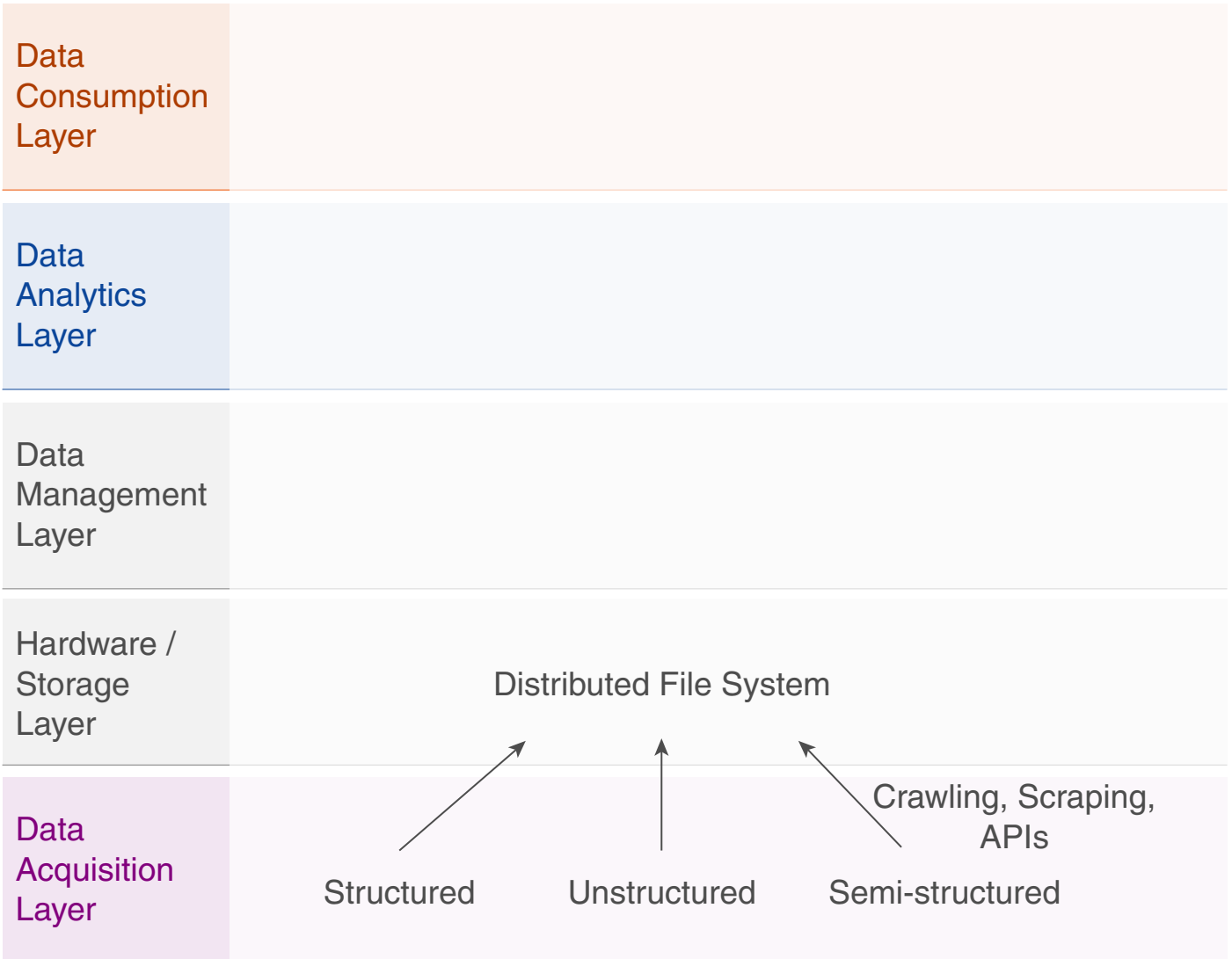
Hardware /
Storage
Layer

Data
Acquisition
Layer

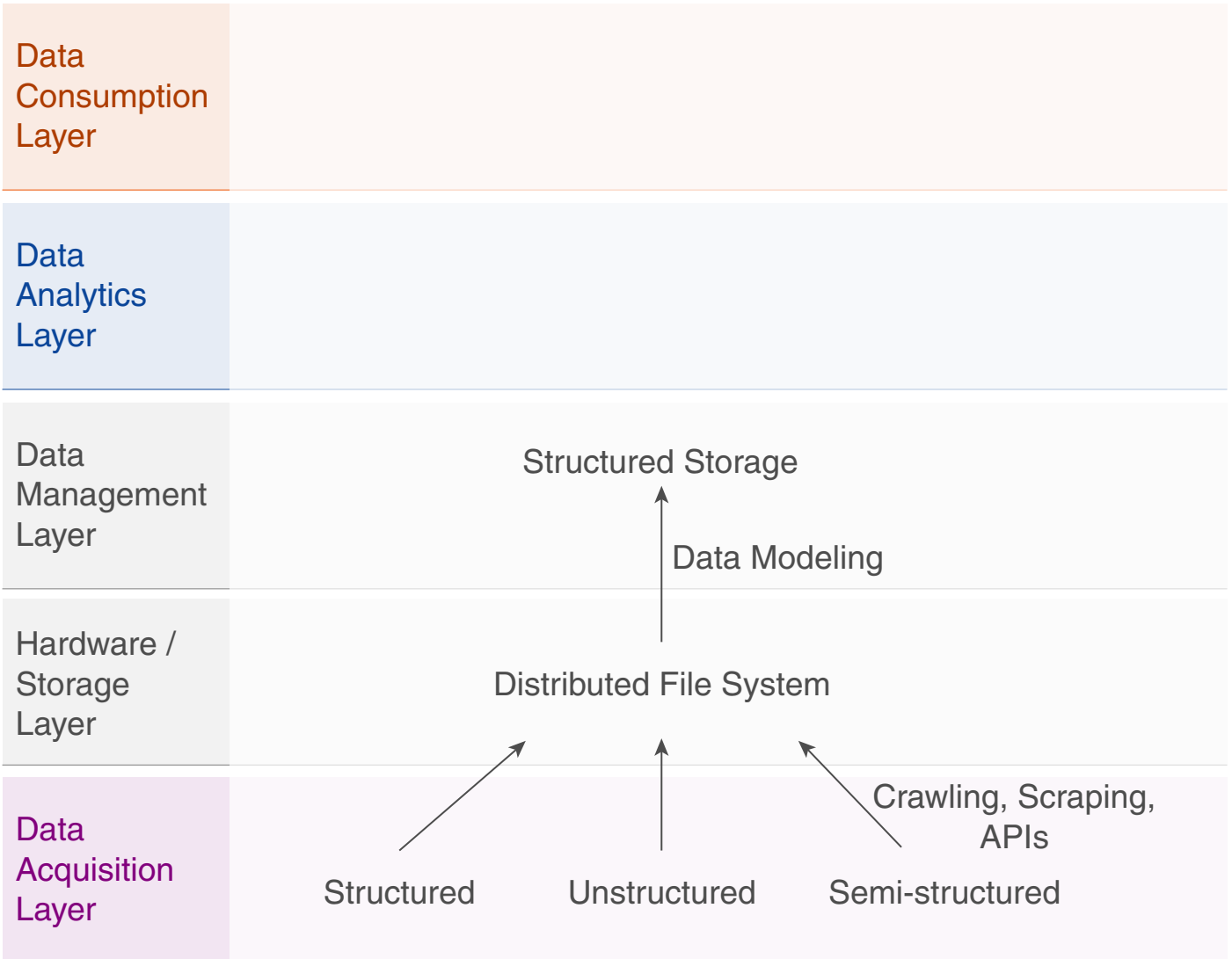
The Big Data Architecture Stack

Data Consumption Layer			
Data Analytics Layer			
Data Management Layer			
Hardware / Storage Layer			
Data Acquisition Layer	Structured	Unstructured	Crawling, Scraping, APIs Semi-structured

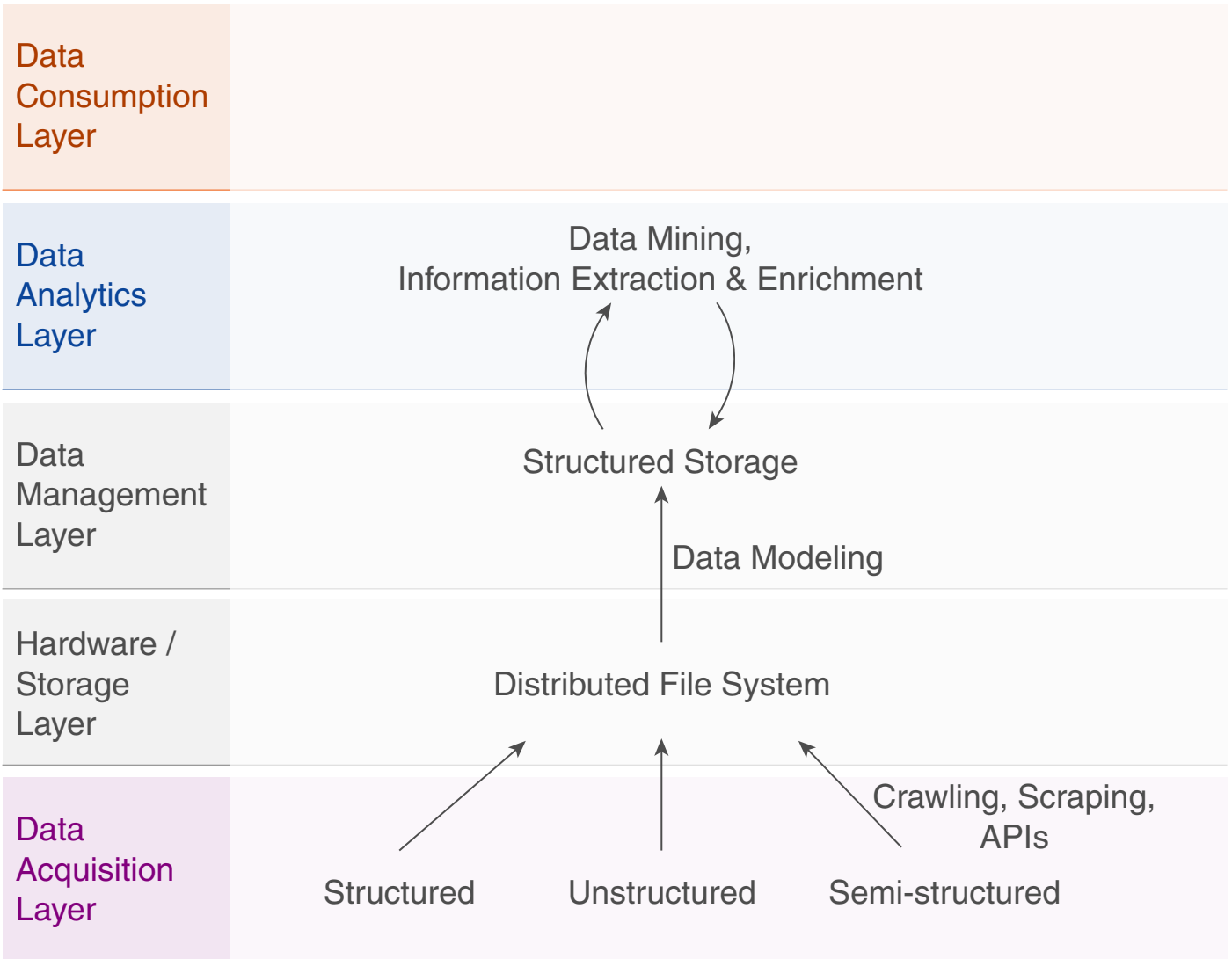
The Big Data Architecture Stack



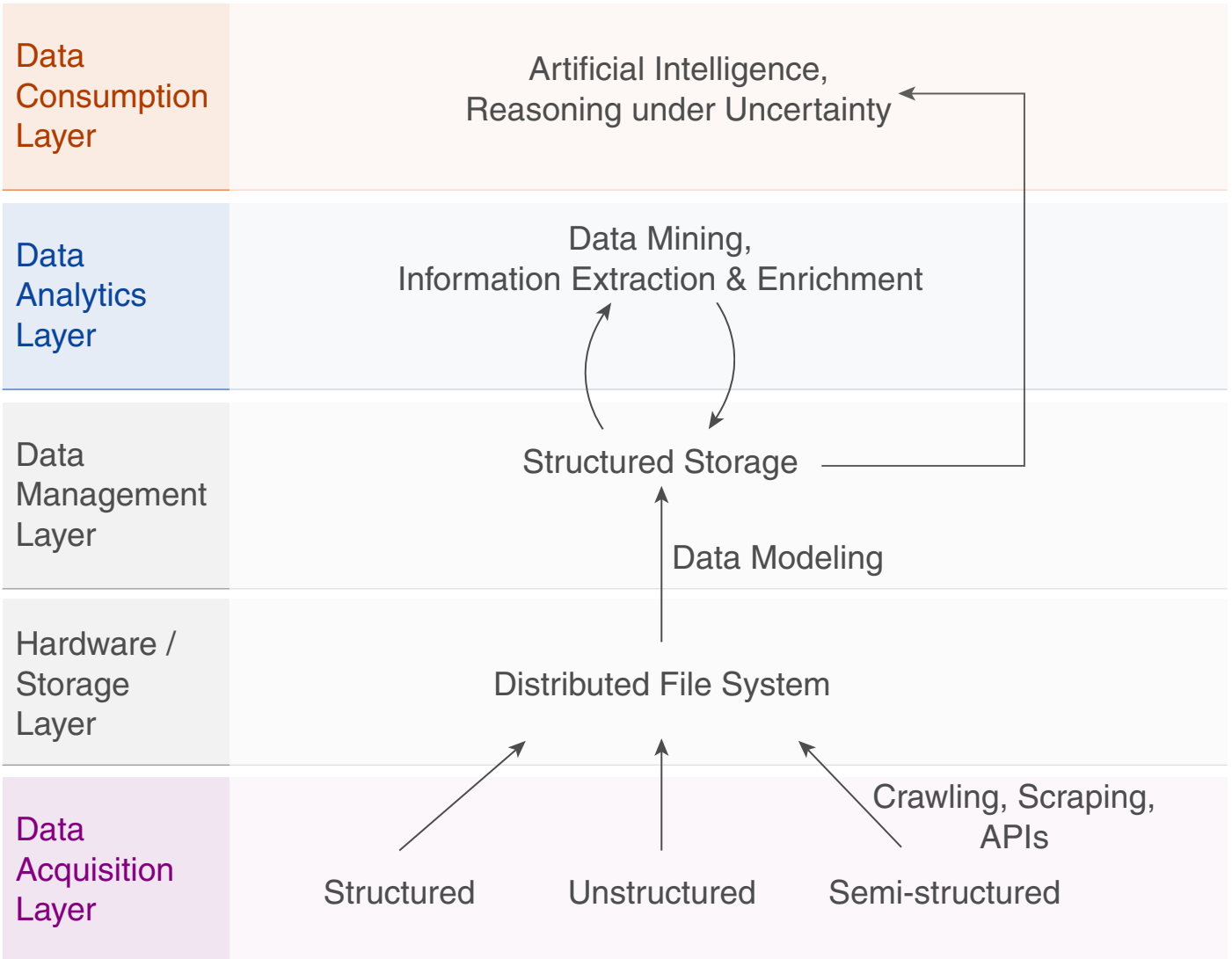
The Big Data Architecture Stack



The Big Data Architecture Stack



The Big Data Architecture Stack

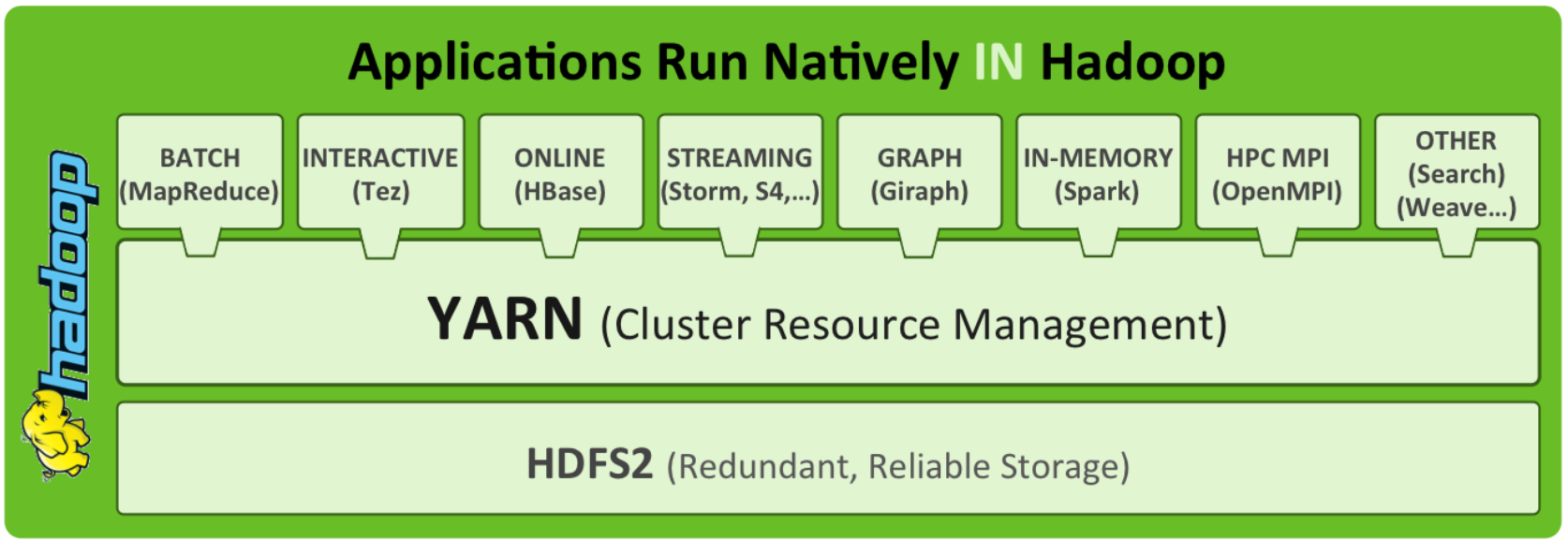


Hadoop YARN

Common Infrastructure for Big Data Technologies

Hadoop YARN

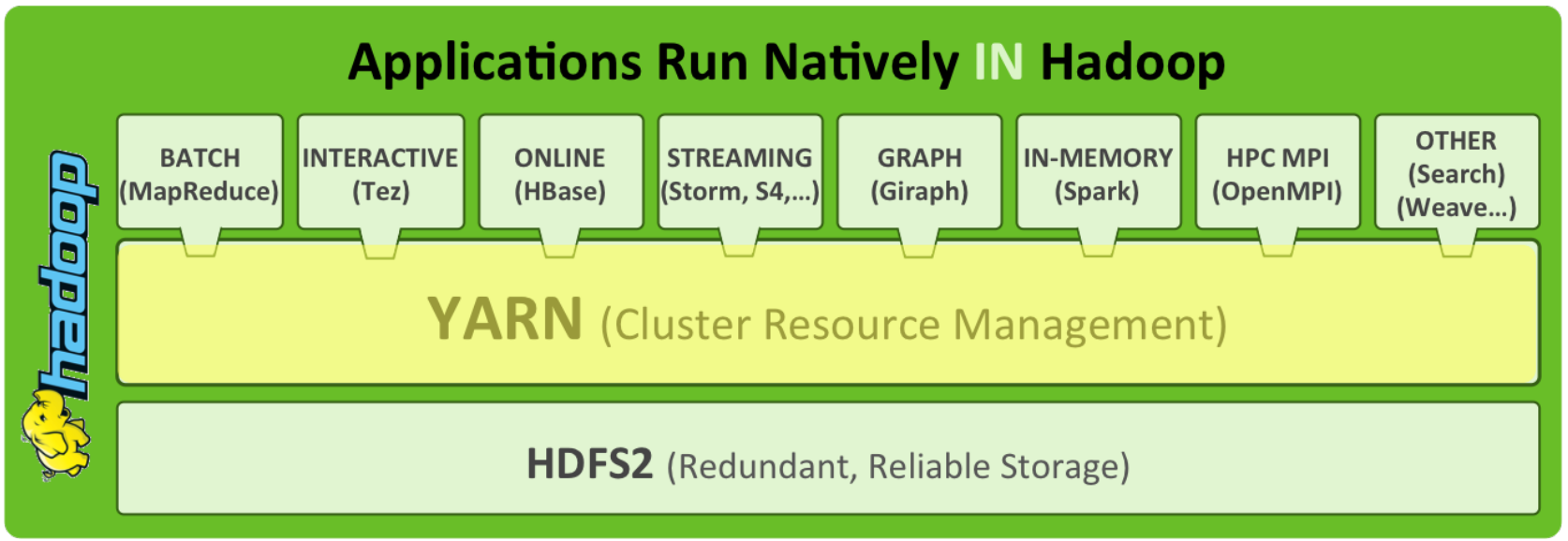
Hadoop 2.0 Ecosystem



[http://hortonworks.com/hadoop/yarn/]

Hadoop YARN

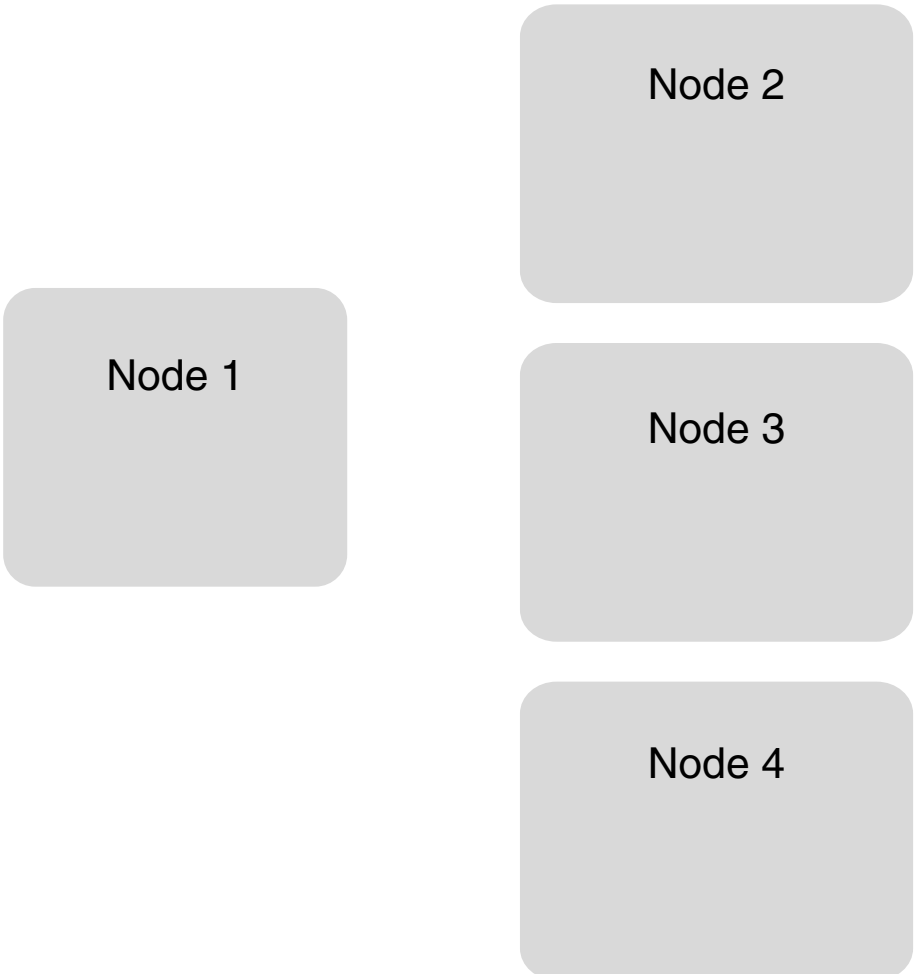
Hadoop 2.0 Ecosystem



[http://hortonworks.com/hadoop/yarn/]

Hadoop YARN

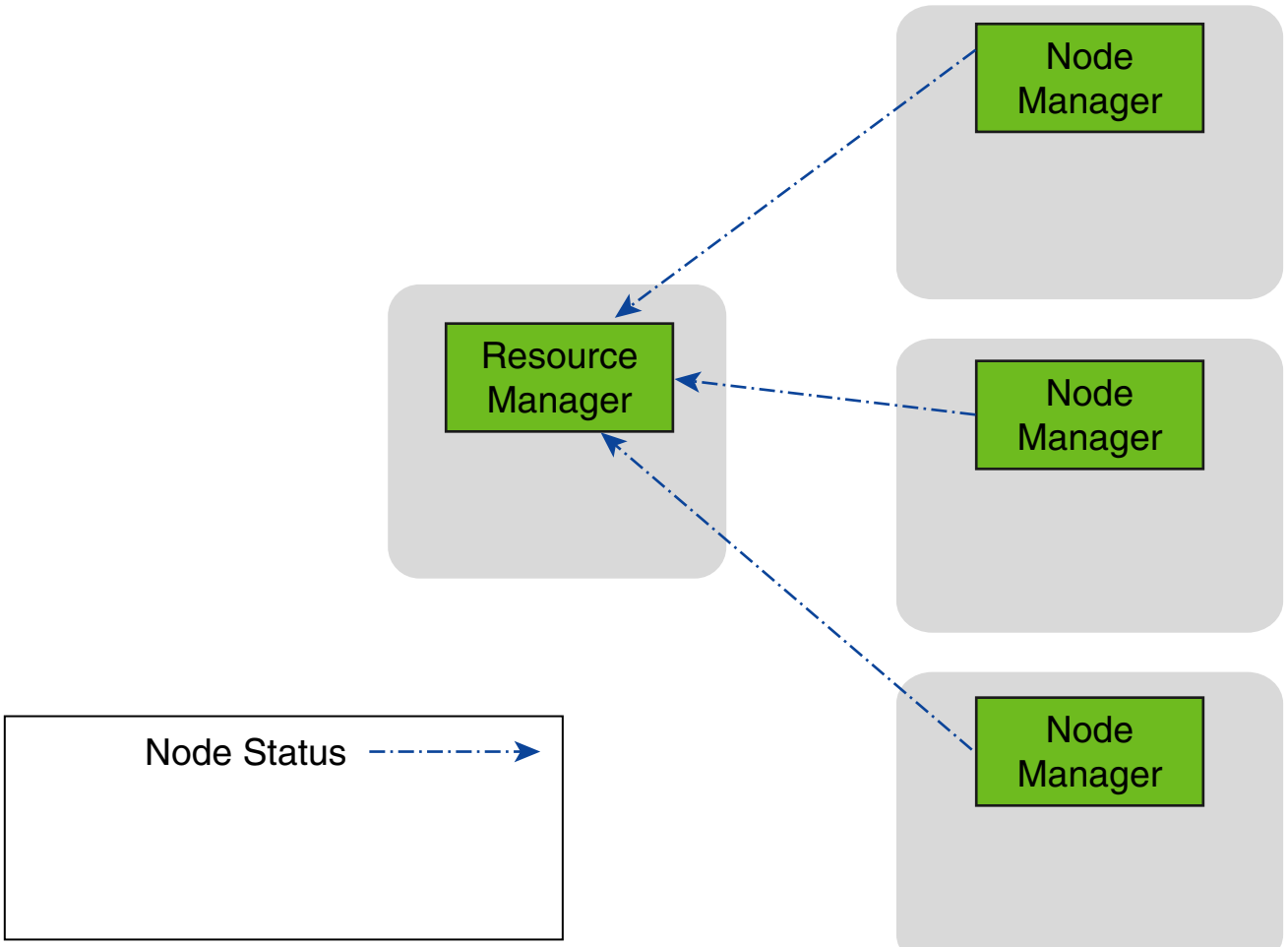
YARN Architecture



[https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html]

Hadoop YARN

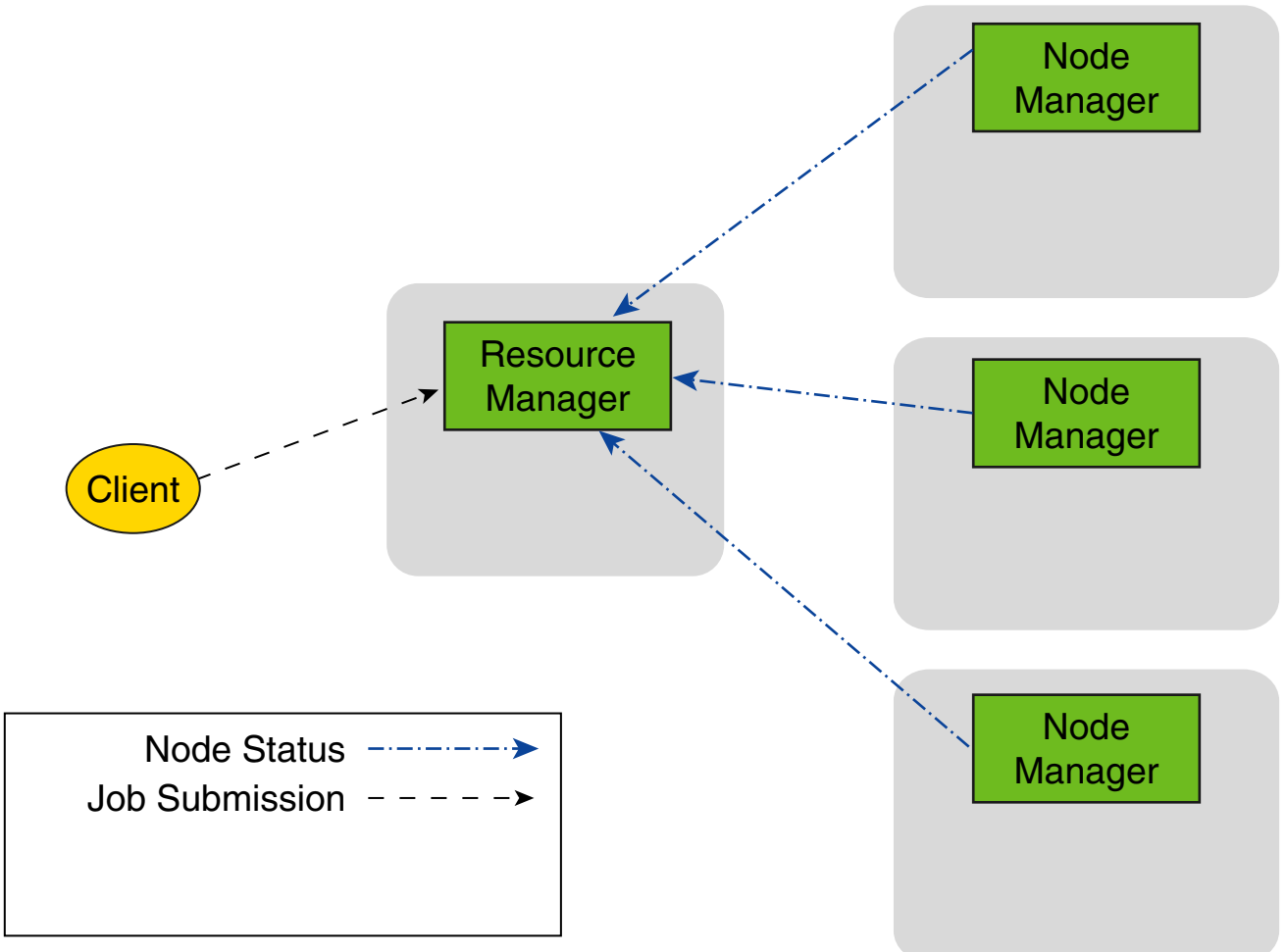
YARN Architecture



[https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html]

Hadoop YARN

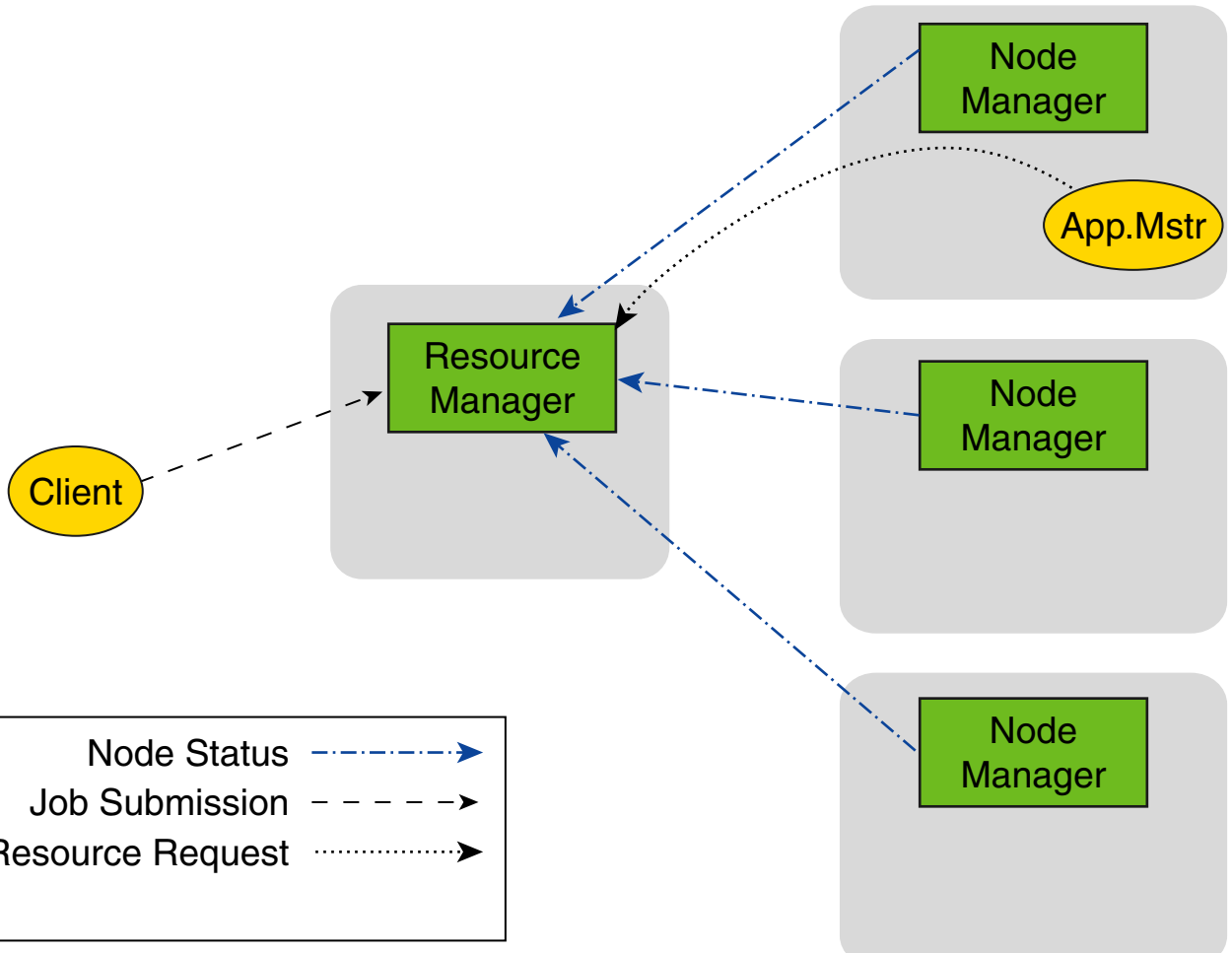
YARN Architecture



[https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html]

Hadoop YARN

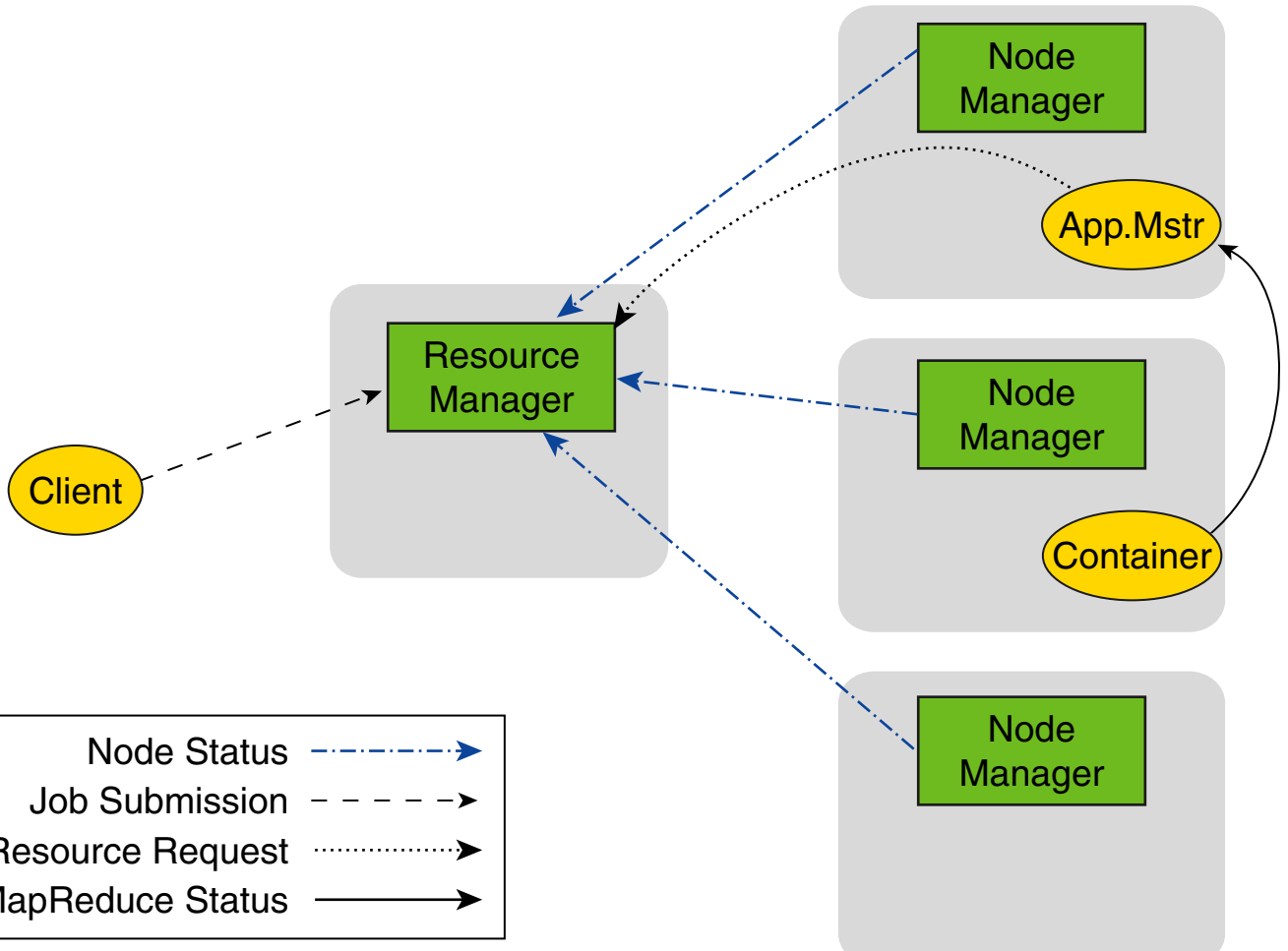
YARN Architecture



[https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html]

Hadoop YARN

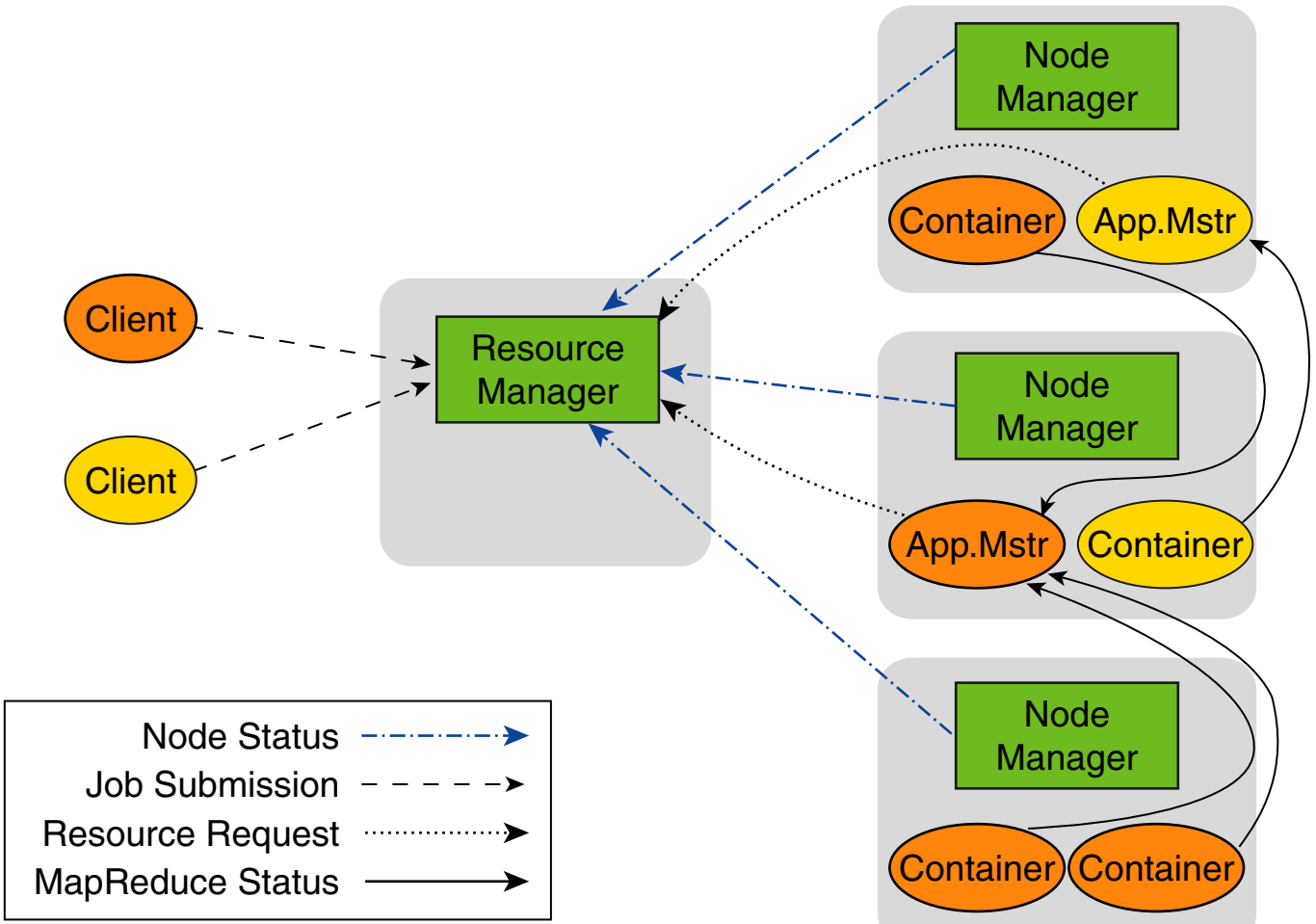
YARN Architecture



[https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html]

Hadoop YARN

YARN Architecture



[https://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html]

betaweb Facts

Our Cluster

- 130 nodes
- 1500 cores
- 8TB RAM
- 2PB HDD

Big Data Architectures For Machine Learning and Data Mining

Seminar Deliverables

1. Short talk
   - 10-20 minutes.
   - Overview of one big data technology.
   - How does it work?
   - What is it good for?
   - Installation instructions.
   - Usage examples.
2. Seminar talk
   - 30 – 45 minutes.
   - Detailed introduction to one big data problem including state-of-the-art.
   - Discussion about possible big data technologies to solve the problem.
   - Presentation of implementation, evaluation, and results.
3. Seminar text
   - High-quality text summarizing findings

Big Data Architectures For Machine Learning and Data Mining

Seminar Topics

Short talk:

- HDFS and basic MapReduce.
- Apache Spark Basics.
- Spark MLLib.
- Spark GraphX.
- Spark Streaming.
- Apache Mahout “Samsara.”
- Apache Flink.
- DeepLearning4J.
- Tensorflow.

Seminar talk (general):

- Near-duplicate detection in large document collections.
- Ad-hoc Search Engine Index.
- Analyzing word similarity with distributed representations.
- Text re-use in Wikipedia.
- Exploring large document collections.
- Social Network Analysis.
- Recommendation Systems.
- Dimensionality Reduction.
- Real-time classification.

Big Data Architectures For Machine Learning and Data Mining

Schedule

This week

- Survey for regular seminar time slot.
  http://doodle.com/poll/9td7qryihy73f295
- Reading:
  Leskovec, Rajaraman, Ullman. *Mining of massive datasets.*
  http://infolab.stanford.edu/~ullman/mmds/book.pdf

Weeks 2-3

- Tutorial:
  Installing Hadoop on virtual machines.
- Preparation of short talks.

Weeks 4-5

- Short Talks.
- Assignment of seminar talk topics.

Dates for the seminar talks are to be determined.

Big Data Architectures For Machine Learning and Data Mining

Thank you!

- Add you name and email address to the participants list.
- Watch the course web page for schedule updates.
  `www.webis.de` → “Teaching” → “SS 2016” → “Big Data Architectures For Machine Learning and Data Mining”
- Homework:
  - Download and install Oracle VirtualBox, as well as the virtual machine image for this course. Links will be provided on the course page.
  - Skim the “Mining of Massive Datasets” book.
  - Take the seminar time slot survey.
  - Further instructions by email.