Big Data Architectures For Machine Learning and Data Mining

Seminar Kick-Off Meeting

May 14, 2018

Web Technology and Information Systems Group

Michael Voelske Shahbaz Syed

<firstname>.<lastname>@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)

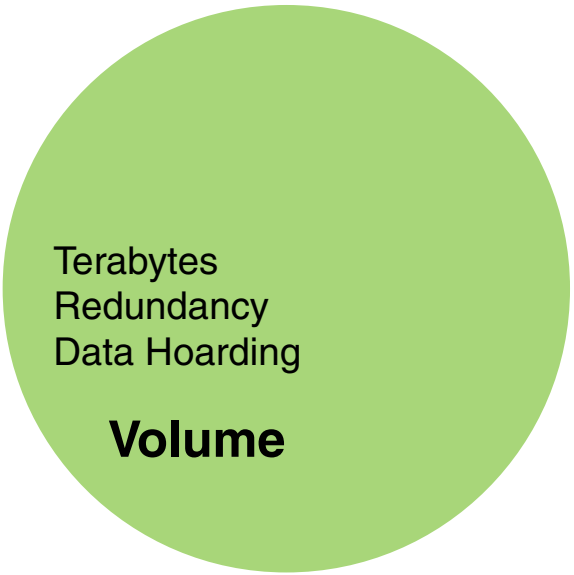
In my view, big data is data that requires novel processing techniques to handle. Typically, **big data requires massive parallelism** in some fashion (storage and/or compute) to deal with volume and processing variety.

Brad Peters (Chief Product Officer, Birst)

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

What is Big Data

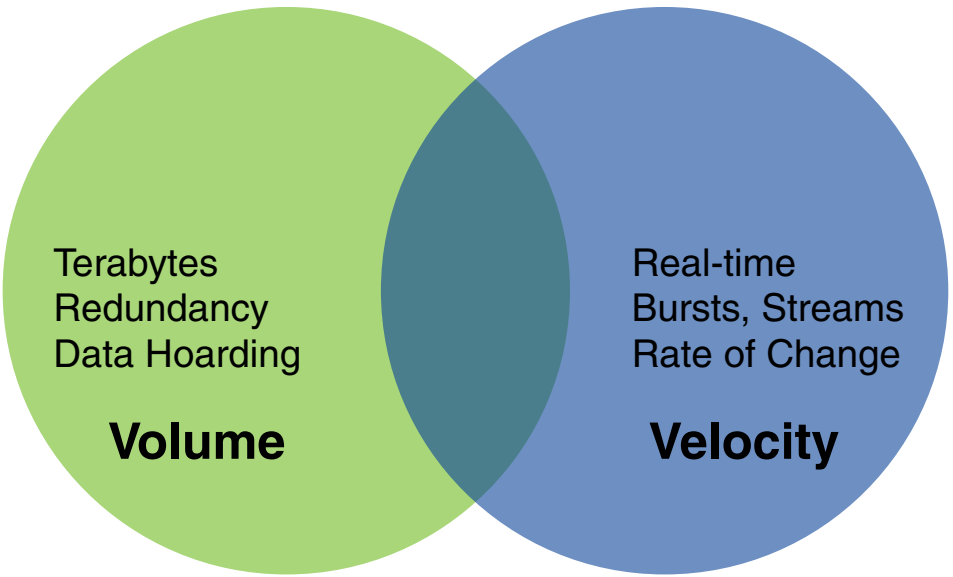
Gartner's "Three V's"



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

What is Big Data

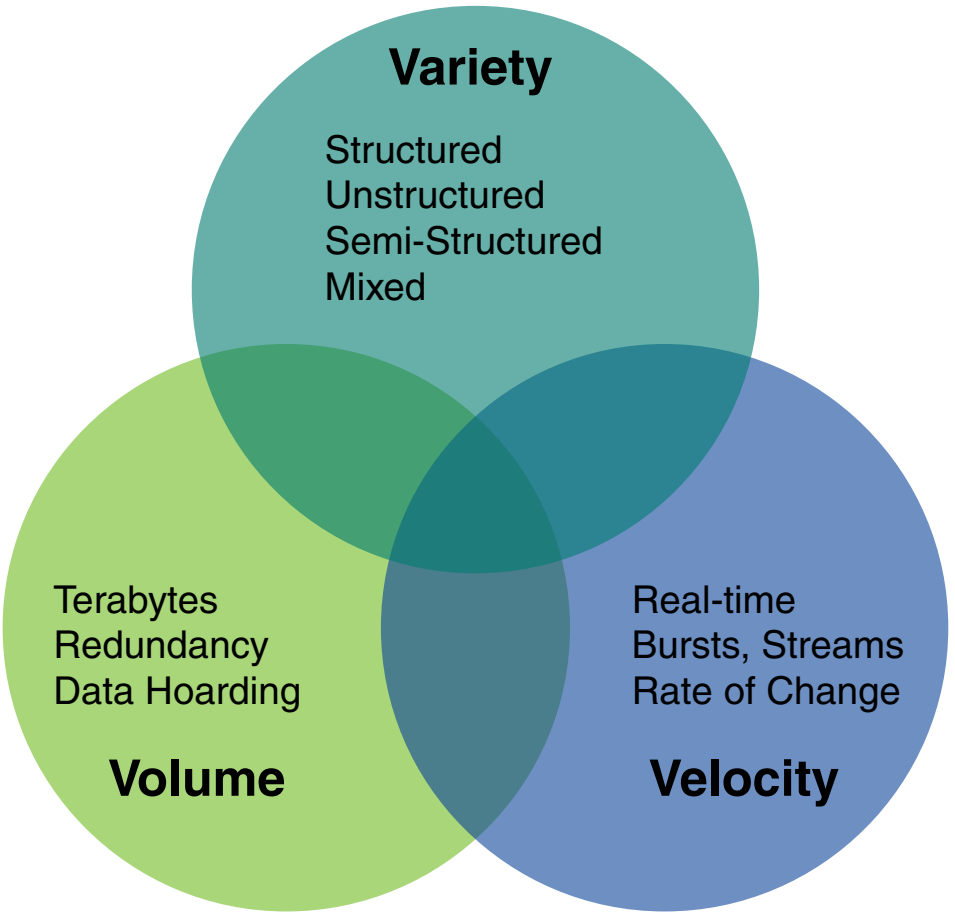
Gartner's "Three V's"



[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

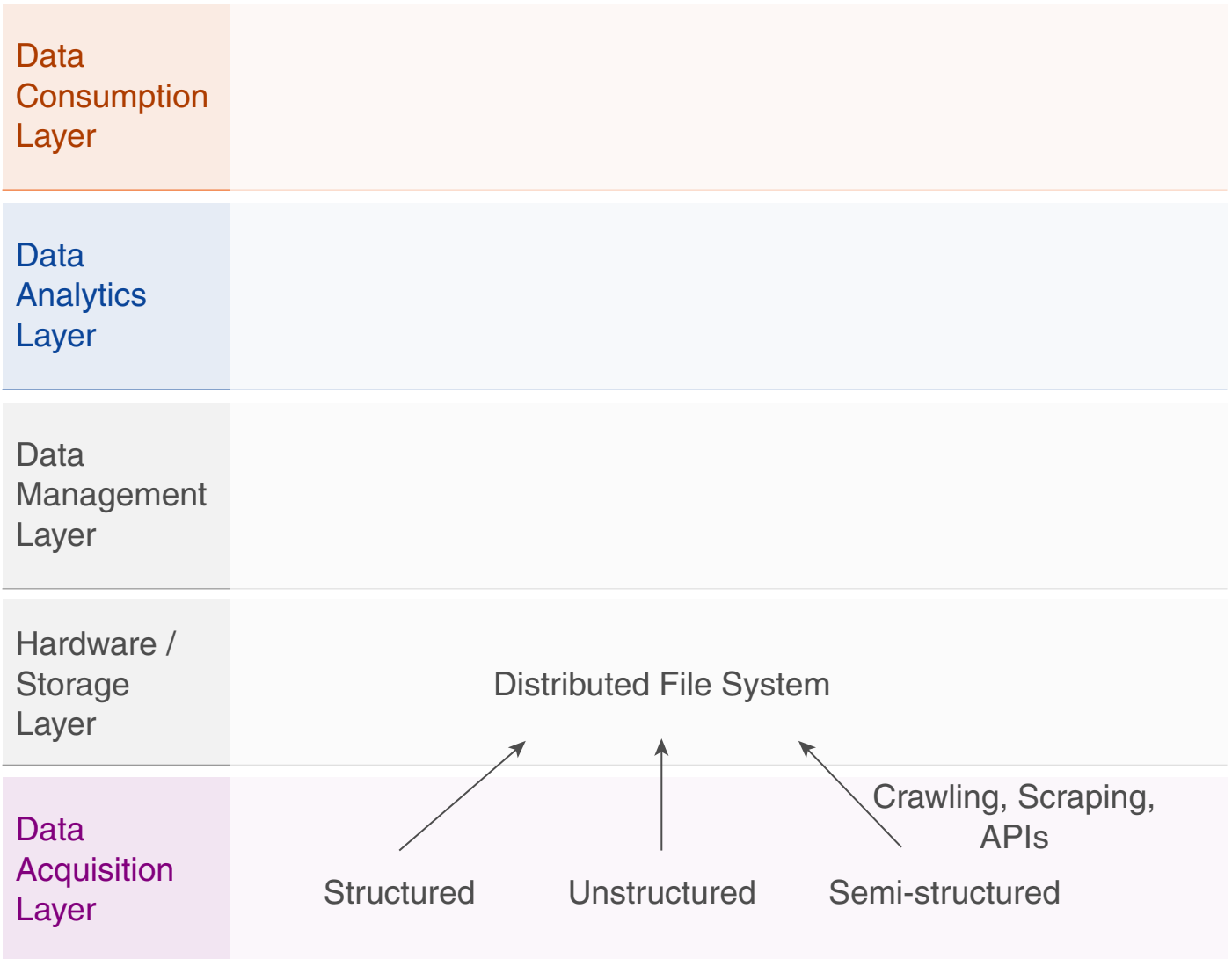
Hardware /
Storage
Layer

Data
Acquisition
Layer

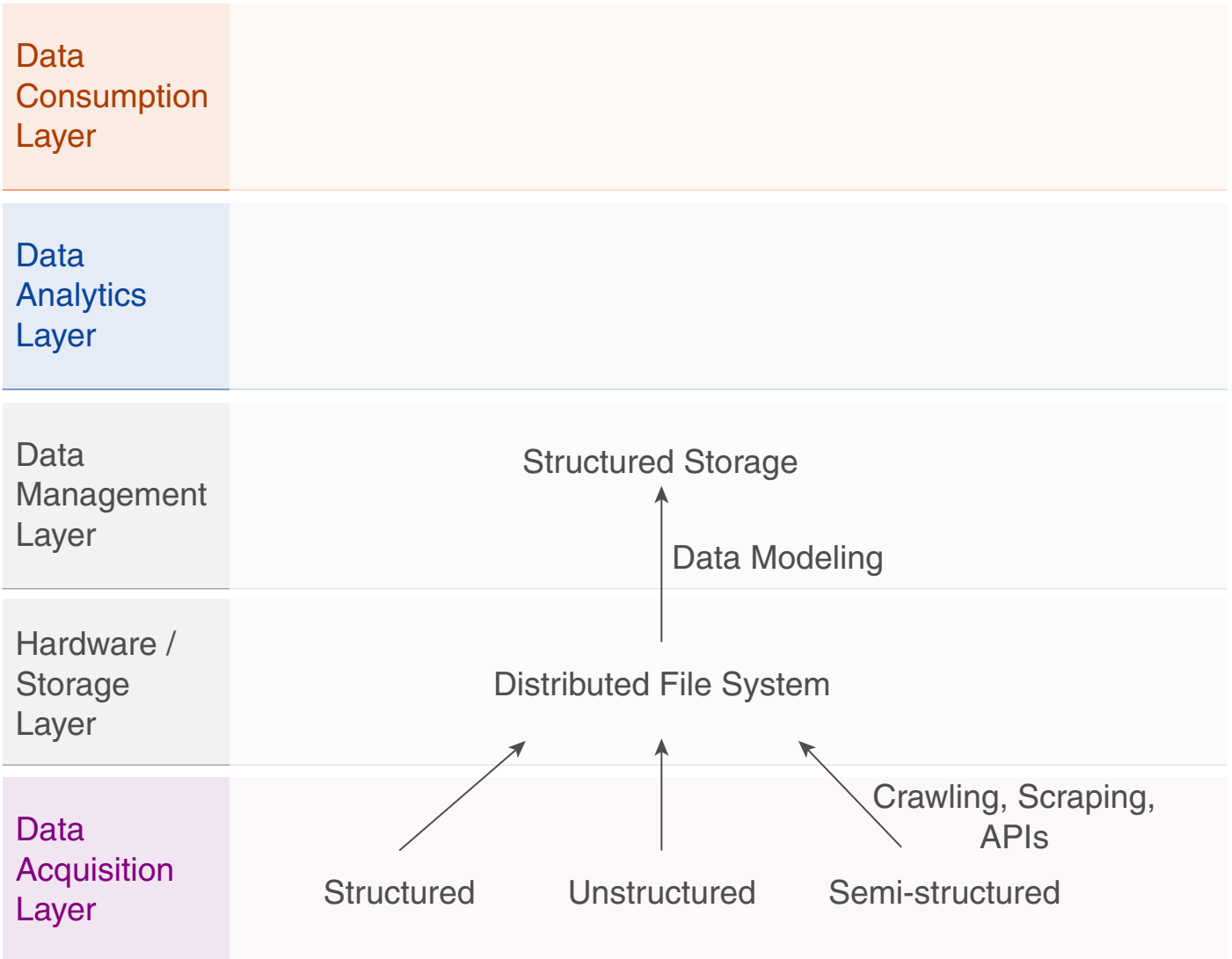
The Big Data Architecture Stack

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

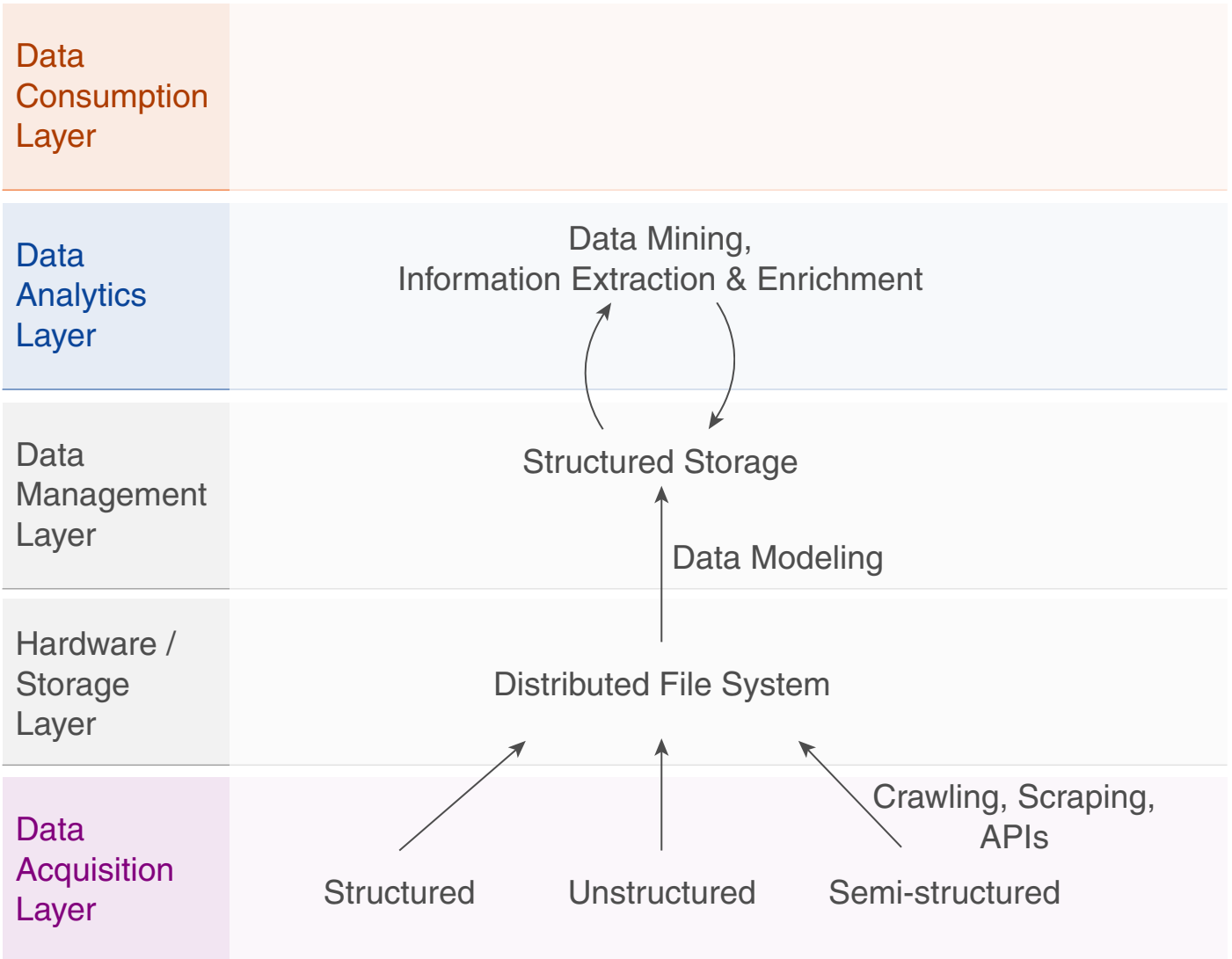
The Big Data Architecture Stack



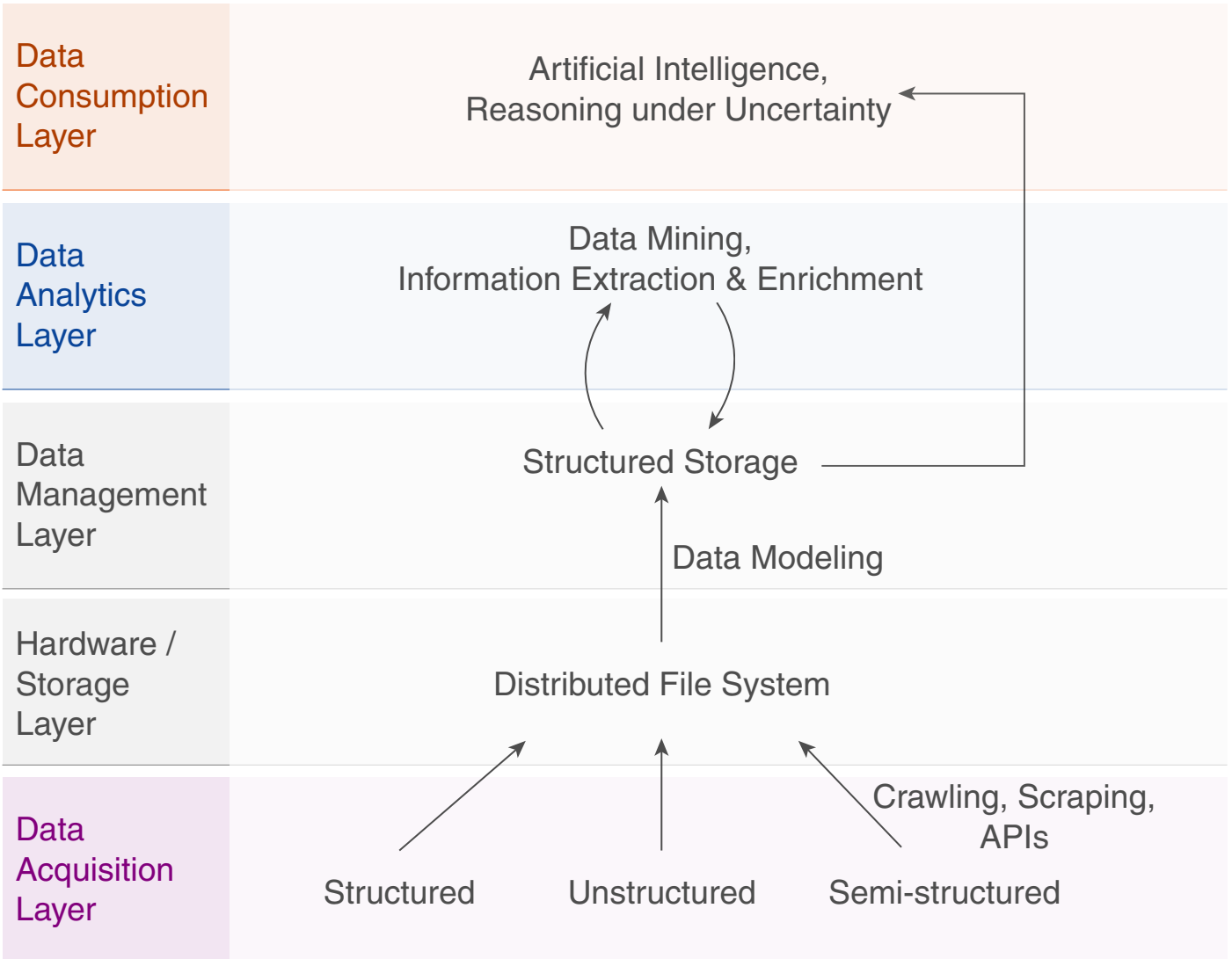
The Big Data Architecture Stack



The Big Data Architecture Stack



The Big Data Architecture Stack

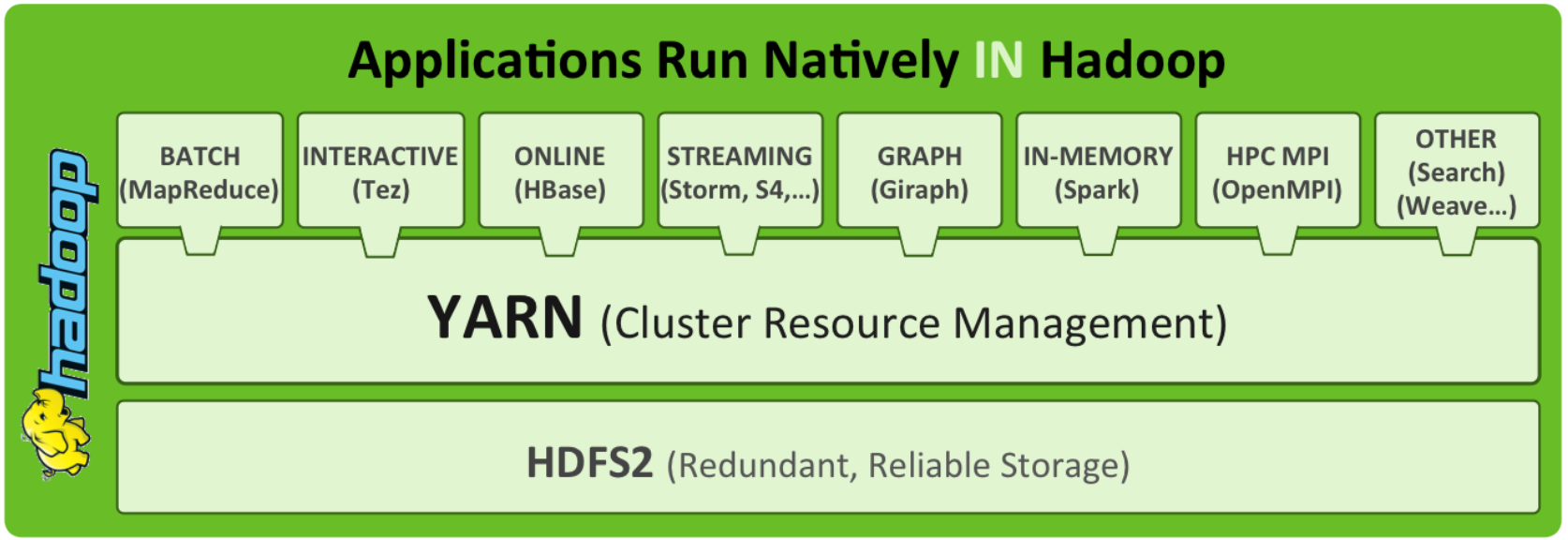


Hadoop YARN

Common Infrastructure for Big Data Technologies

Hadoop YARN

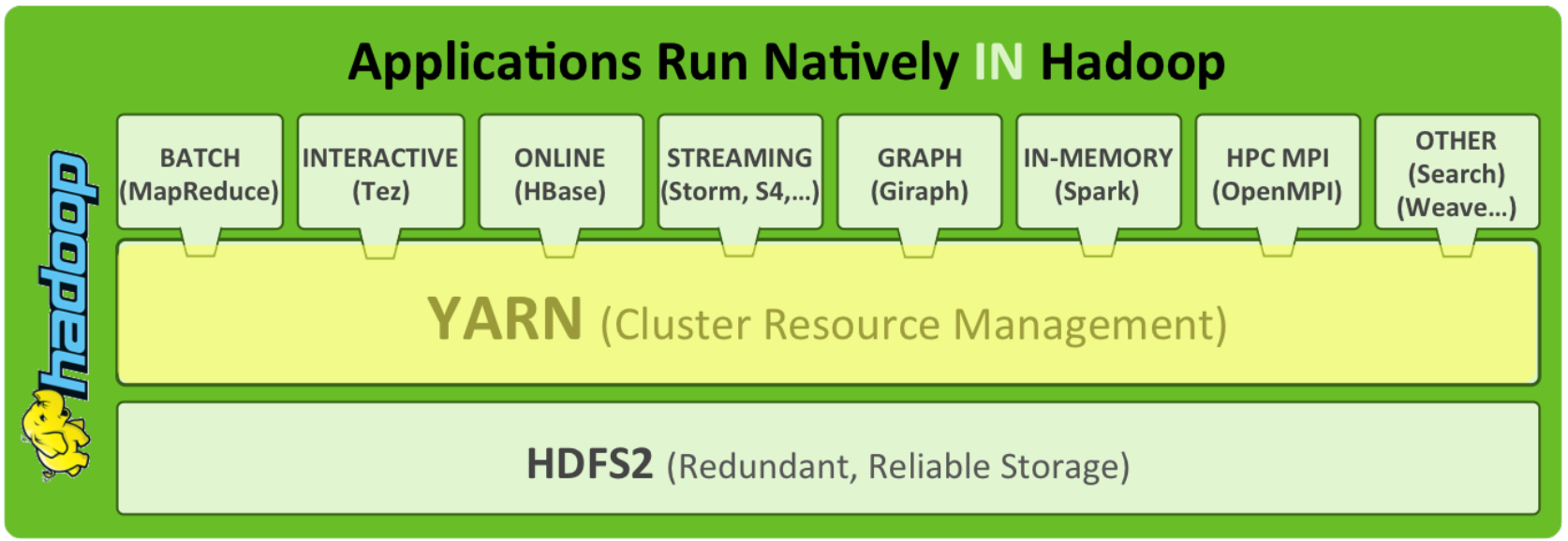
Hadoop 2.0 Ecosystem



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

Hadoop YARN

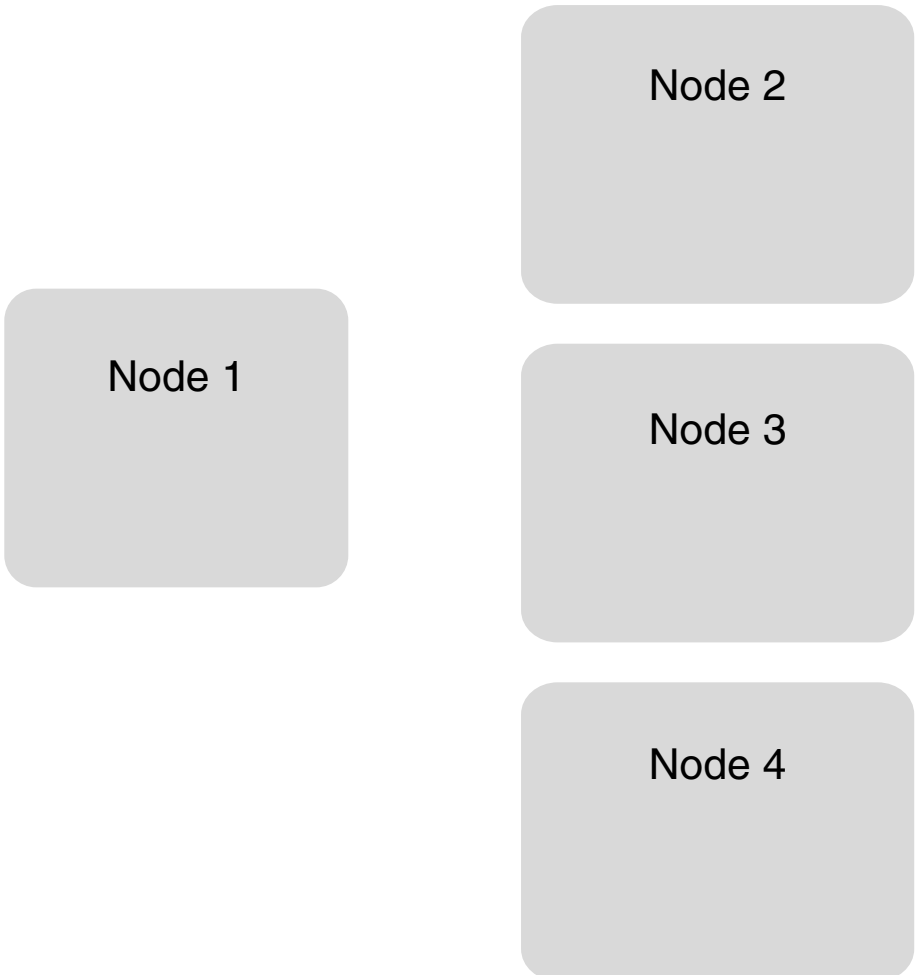
Hadoop 2.0 Ecosystem



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

Hadoop YARN

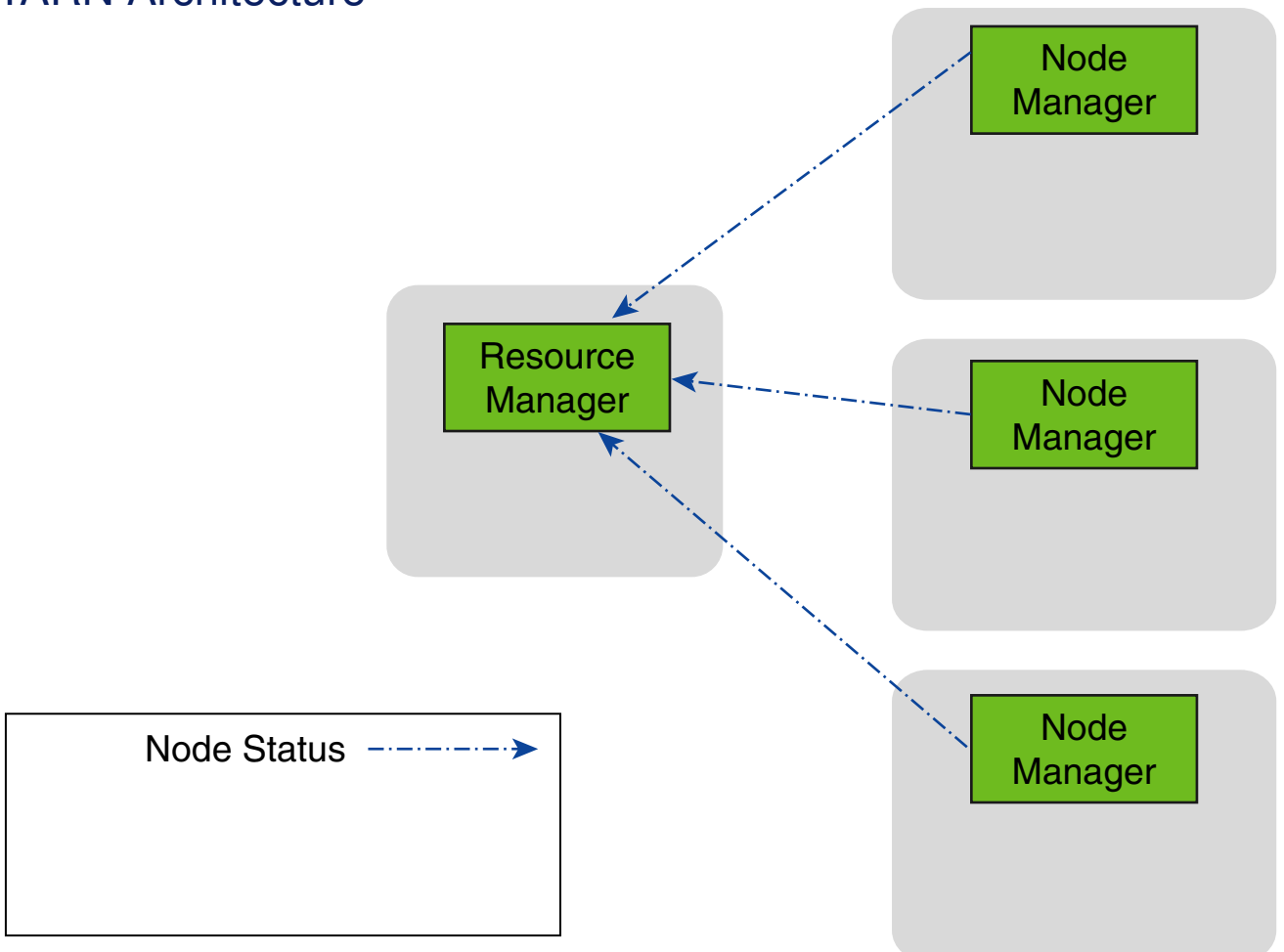
YARN Architecture



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

Hadoop YARN

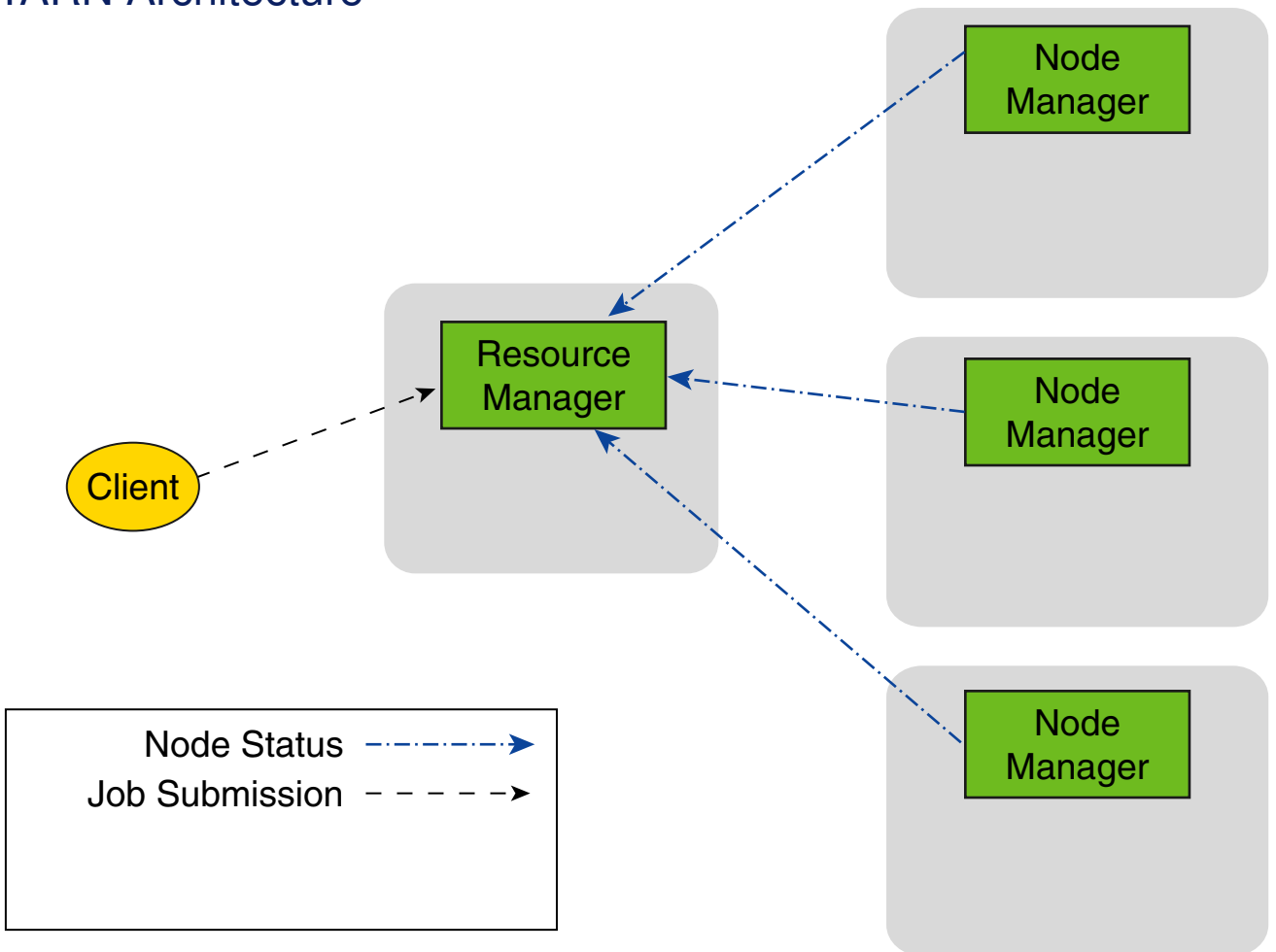
YARN Architecture



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

Hadoop YARN

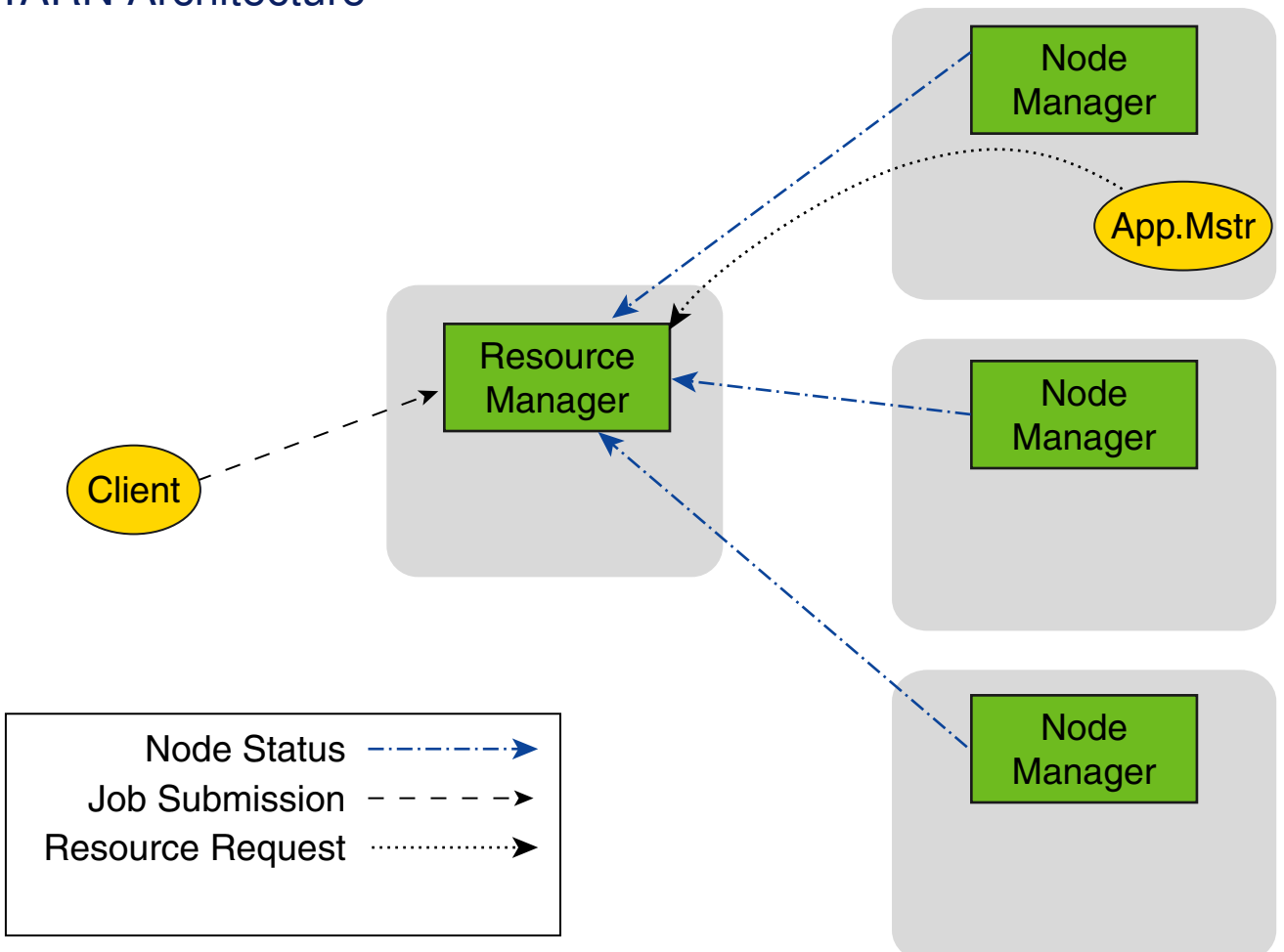
YARN Architecture



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

Hadoop YARN

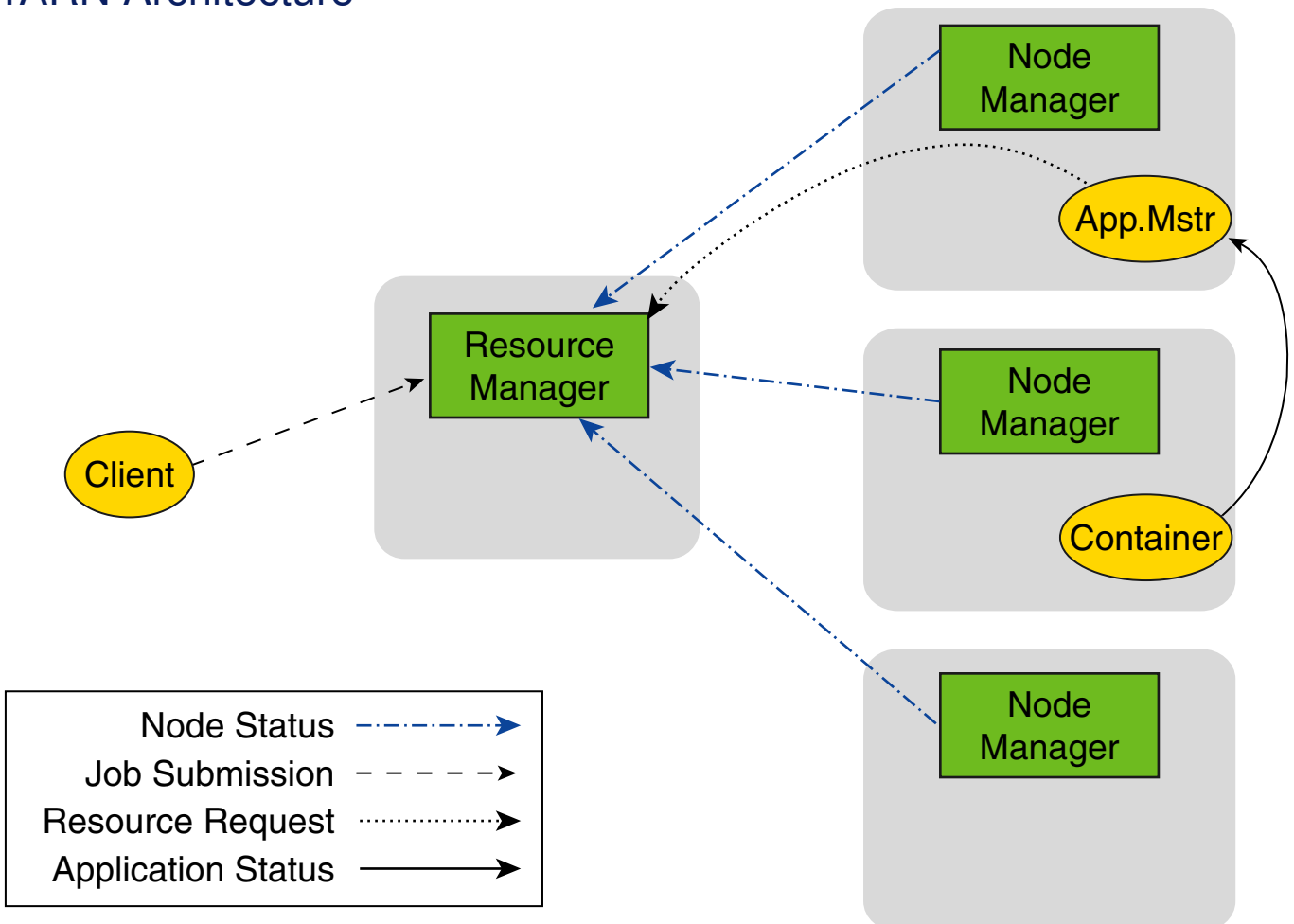
YARN Architecture



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

Hadoop YARN

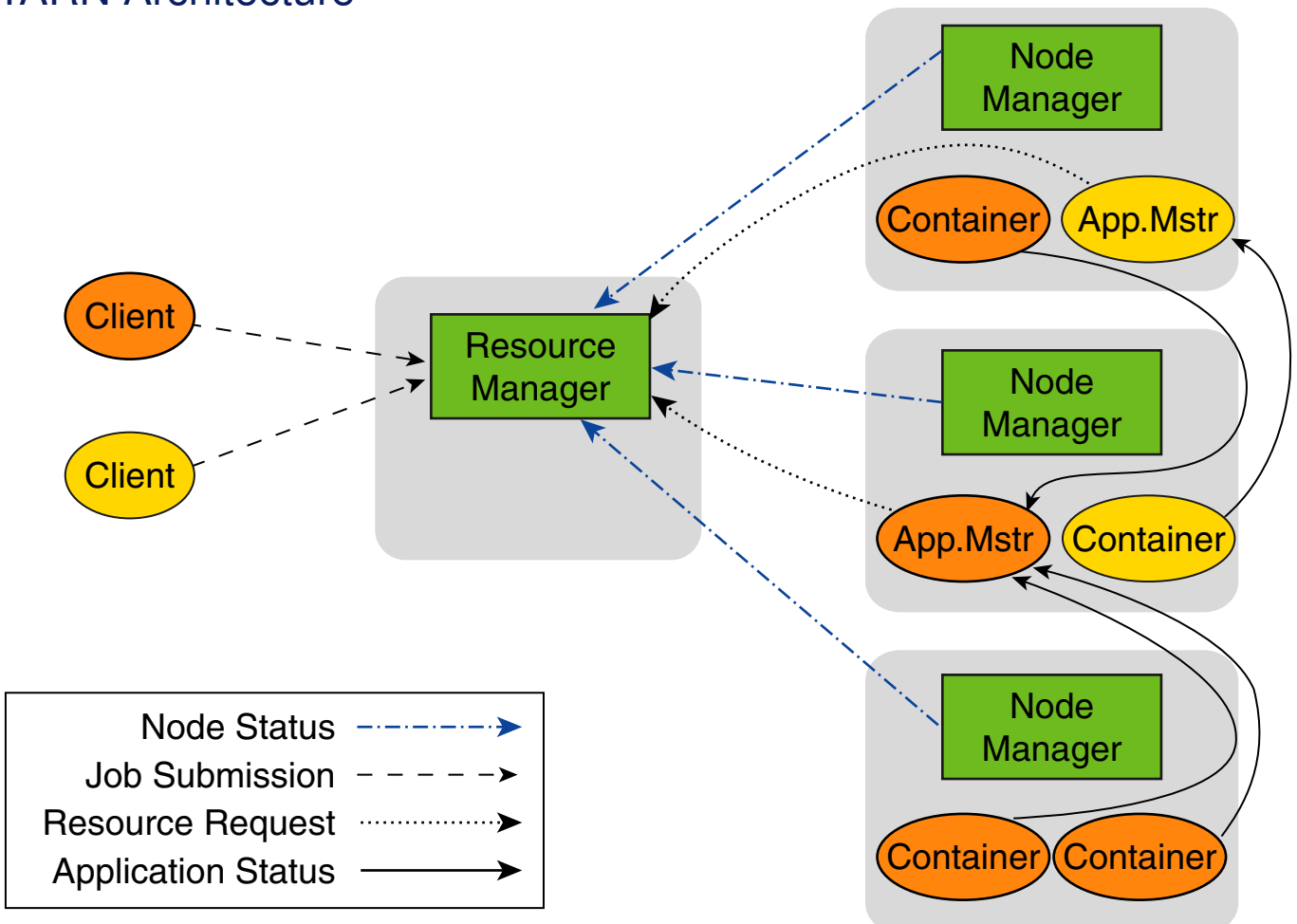
YARN Architecture



[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
- 24TB RAM
- 2PB HDD

Big Data Architectures For Machine Learning and Data Mining

Seminar Deliverables

1. Short talk
   - 10-15 minutes.
   - Overview of one big data/ML technology: What / How / Why?
   - Installation instructions.
   - Usage examples.
   - List of topics will be provided. Approach us with own ideas!
2. Seminar talk
   - 30 – 45 minutes.
   - Solve one big data problem given a particular dataset.
   - Discuss problem-solving approach, issues, state-of-the-art.
   - Present implementation, evaluation, and results.
   - List of topics will be provided. Approach us with own ideas!
3. Seminar paper
   - 4 – 8 pages.

- High-quality text summarizing findings (same topic as seminar talk)

Big Data Architectures For Machine Learning and Data Mining

Schedule

This week

- Reading:
  Leskovec, Rajaraman, Ullman. *Mining of massive datasets.*
  http://infolab.stanford.edu/~ullman/mmds/book.pdf

Weeks 2-3

- Tutorial:
  Getting started with Hadoop and Docker.
- 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 your name and email address to the participants list.
- Watch the course web page for schedule updates.
  `www.webis.de` → “Teaching” → “SS 2018” → “Big Data Architectures For Machine Learning and Data Mining”
- Homework:
  - Download and install Docker CE. Links will be provided on the course page.
  - Skim the “Mining of Massive Datasets” book.
  - Further instructions by email.

Big Data Architectures For Machine Learning and Data Mining

Short talks

1. **YARN: Job scheduling** *Available scheduling algorithms, configuration, fault-tolerance*
2. **HDFS: Java-based distributed file system** *Namenode & Datanode internals, Data replication & organization, Replication vs erasure codes (HDFS3)*
3. **Apache Spark: Cluster computing framework for Big data** *Introduction to Spark, Installation & tuning parameters, Configuring a Spark application, Local vs Cluster mode for yarn, Application monitoring UI*
4. **Spark RDD: Partitioned collections for parallel processing** *What is it, how does it work, what can it do?*
5. **Apache Hive: SQL-based data warehouse for Big data** *Database vs Data warehouse, Architecture, Views and Indexes*

Big Data Architectures For Machine Learning and Data Mining

Short talks

6. **Spark SQL & Dataframes** *Dataframes, Dataframe vs RDD, Spark SQL, Creating and operating dataframes with examples in spark-shell*
7. **Parquet: Compressed, columnar file format** *Column-oriented DBMS, Why do we need Parquet?, File format, Performance comparison with JSON format*
8. **Introduction to Spark.MLlib** *Basic Statistics, Pipelines and Feature Extraction/Selection/Transformation*
9. **Classification with Spark.MLlib** *Define Classification, Logistic Regression & Naive Bayes*
10. **Regression with Spark.MLlib** *Define Regression, Linear Regression & Random Forest*

Long talk topics

Seminar Projects Overview

- Select a topic from the following three
- Work in groups of 3-4 students
- Each topic will be given a corresponding dataset, available in our HDFS cluster
- Explore and analyse the data, then develop and address a small research problem
- Present your results in a seminar talk and seminar paper and the end of the semester

Long talk topics

1. Examining user review quality for predicting usefulness

- Dataset : Amazon Reviews
- Collect all reviews for atleast 2 categories
- Analyze word distributions and correlation with review score
- Try to answer the question: what makes a review helpful?


```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the
piano.  He is having a wonderful time playing these old hymns.
The music  is at times hard to read because we think the book
was published for singing from more than playing from.  Great
purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```


3 Students

Long talk topics

2. Meta analysis of digital news content

- Dataset : GDELT 2.0 (Global Database of Events, Language and Tone)
- Explorative analysis of a large collection of news event metadata (billions of events)
- Extract more focused sub-datasets
- Devise interesting visualizations

GDELT 2.0 - Components

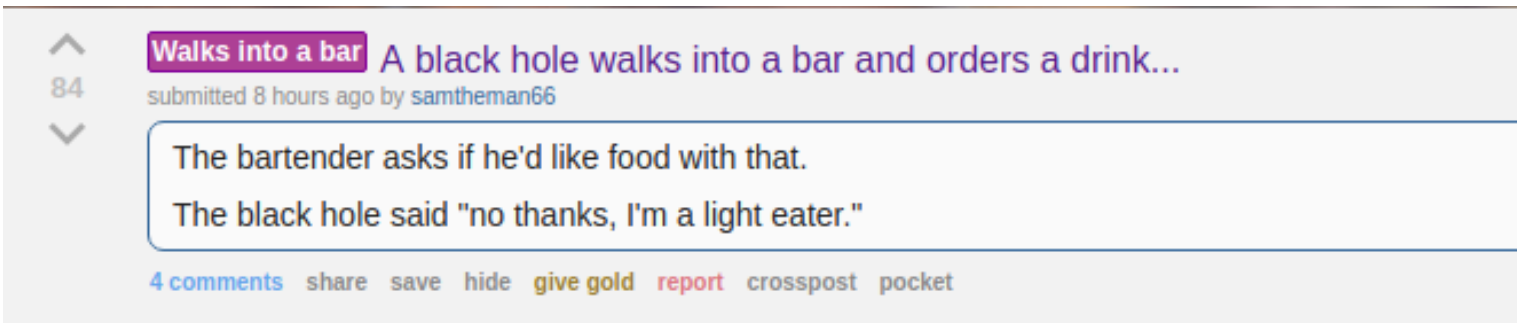
1. Events : All news events across the globe are analysed and multiple features are extracted
2. Global Knowledge Graph (GKG) : Knowledge graph of the events database connecting all pariticipating entities, types of events, emotions, and themes for each news event
3. Mentions : Records each mention of an event for tracking the trajectory and network structure of a story as it flows through the global media system

4 Students

Long talk topics

3. Is this a joke ?

- Dataset : Reddit (r/jokes)
- Extract English jokes with their scores from a large social media site
- Design a simple regression model that given a joke, predicts its score
- Try to answer the question: what makes a joke funny?



3 Students

Outlook

Organization

- Today: select topics and groups.
- Initial reading material and a data sample in the next few days.
- Access to betaweb cluster and full data soon; tutorial in 2 weeks.
- Work on the project until July 2nd (Long talk presentations).
- Weekly meetings (Mon 11:00 or other appointment).
- Seminar paper due August 10th (two weeks after the exam period).

Long Talks

- About 30 minutes
- Present the problem with background and related work
- Overview and basic statistics of the dataset
- Your approach and results
- Key lessons learned in the implementation