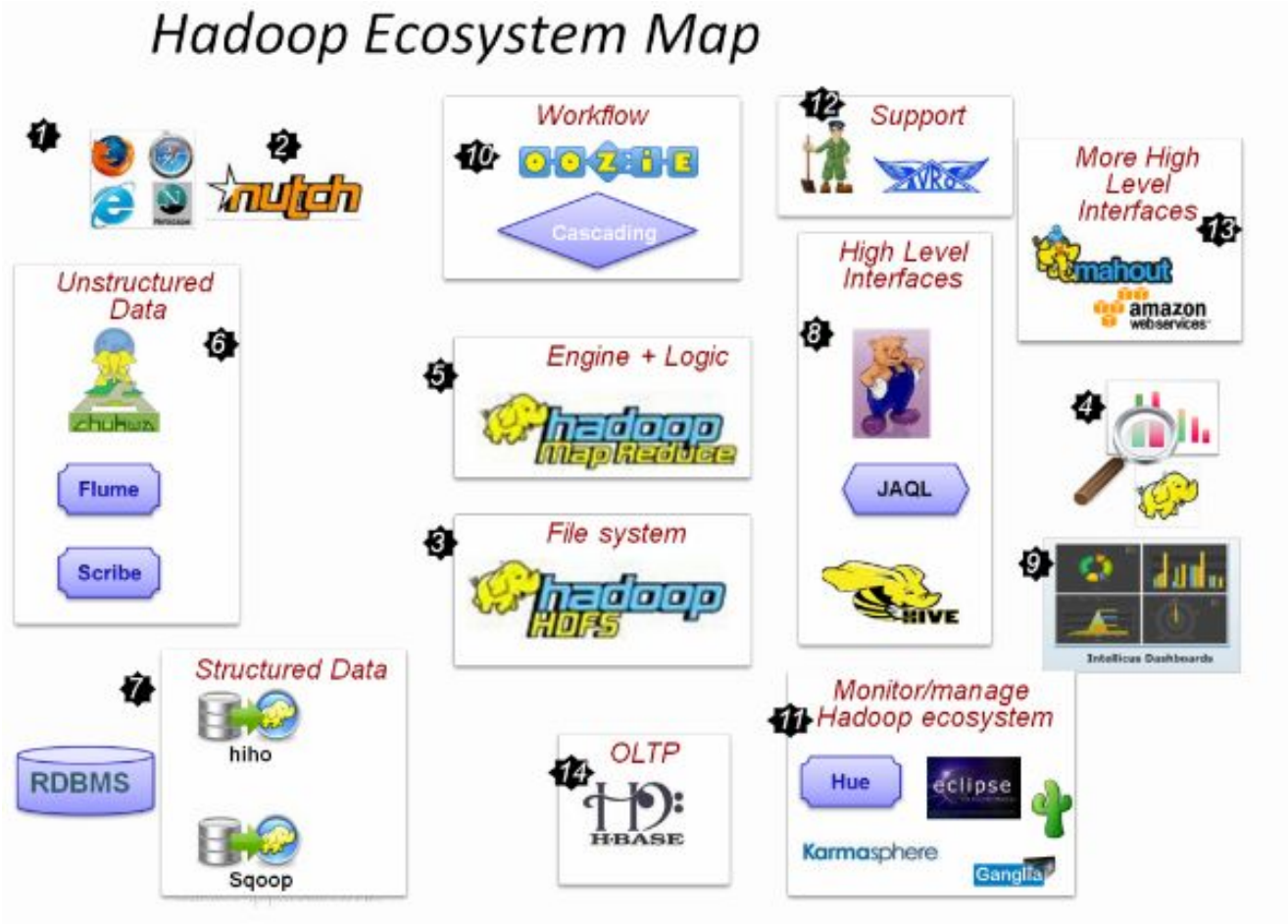
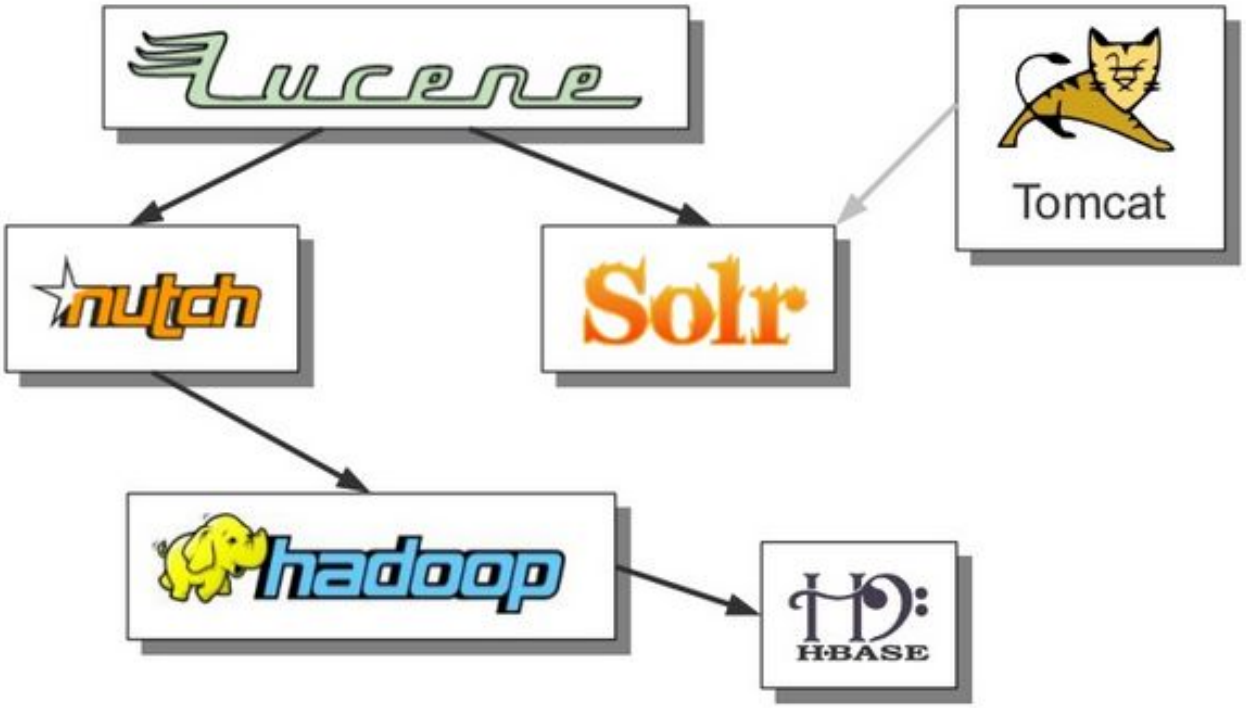
HADOOP

- Григорьев А., ТИ-51м


Hadoop Ecosystem Map
1
2
nutch
Workflow
10
OOZIE
Cascading
12
Support
More High Level Interfaces
13
mahout
amazon webservices
Unstructured Data
6
zhuhaizi
Flume
Scribe
High Level Interfaces
8
JAQL
HIVE
Engine + Logic
5
hadoop Map Reduce
4
File system
3
hadoop HDFS
9
Intellicus Dashboards
Structured Data
7
hiho
RDBMS
Sqoop
Monitor/manage Hadoop ecosystem
11
Hue
eclipse
Karmasphere
Ganglia
OLTP
14
HBASE

Кто использует Hadoop

История развития проектов Apache Software Foundation



Принципы Hadoop

Горизонтальное масштабирование вместо вертикального

- Сложнее и дороже масштабироваться “вверх”
  - Добавить дополнительные ресурсы к существующему железу (CPU, RAM)
  - Закон Мура не успевает за ростом объема данных
  - Если нельзя улучшить железо, то надо покупать более мощное новое
  - Это вертикальное масштабирование
- Горизонтальное масштабирование
  - Добавить больше машин к существующему распределенному окружению
  - Уровень приложения поддерживает добавление/удаление нод
  - Hadoop исповедует такой подход – набор связанных нод
  - Так же очень просто масштабироваться “вниз”

Принципы Hadoop

Отказы оборудования

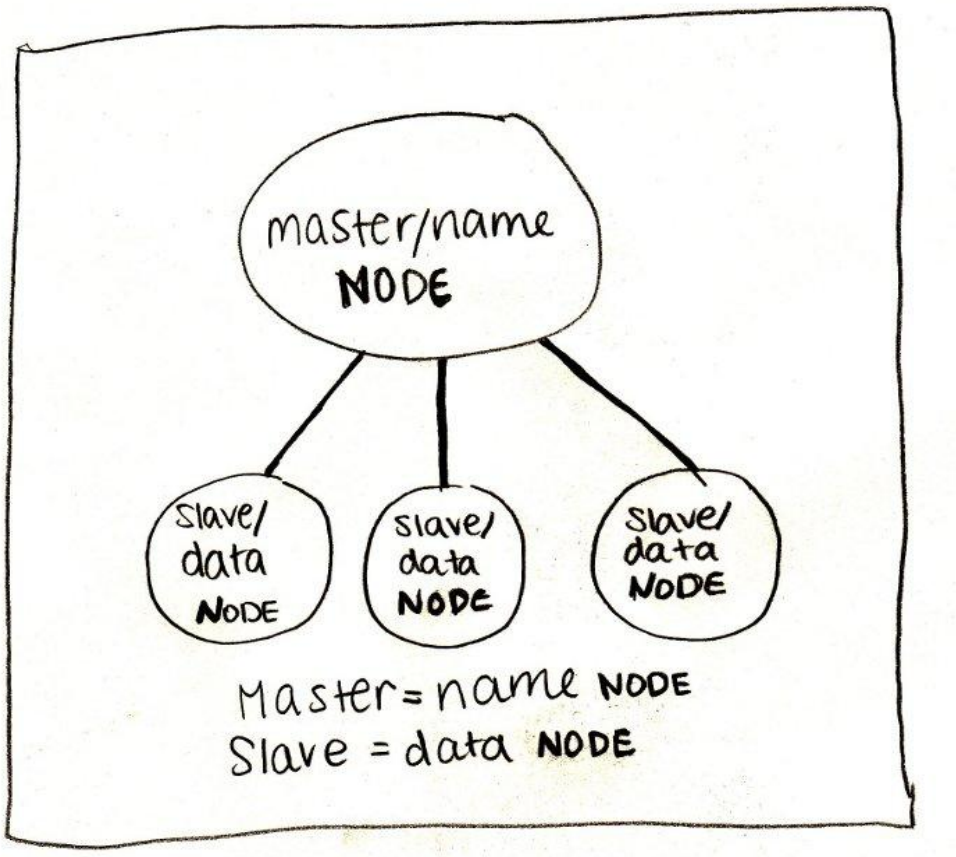
- Чем больше количество машин, тем чаще будут отказы железа
  - На больших кластерах (сотни и тысячи машины) отказы будут еженедельно (и даже ежедневно!)
- Hadoop разрабатывался с учетом отказов железа
  - Репликация данных
  - Перезапуск тасков

Принципы Hadoop

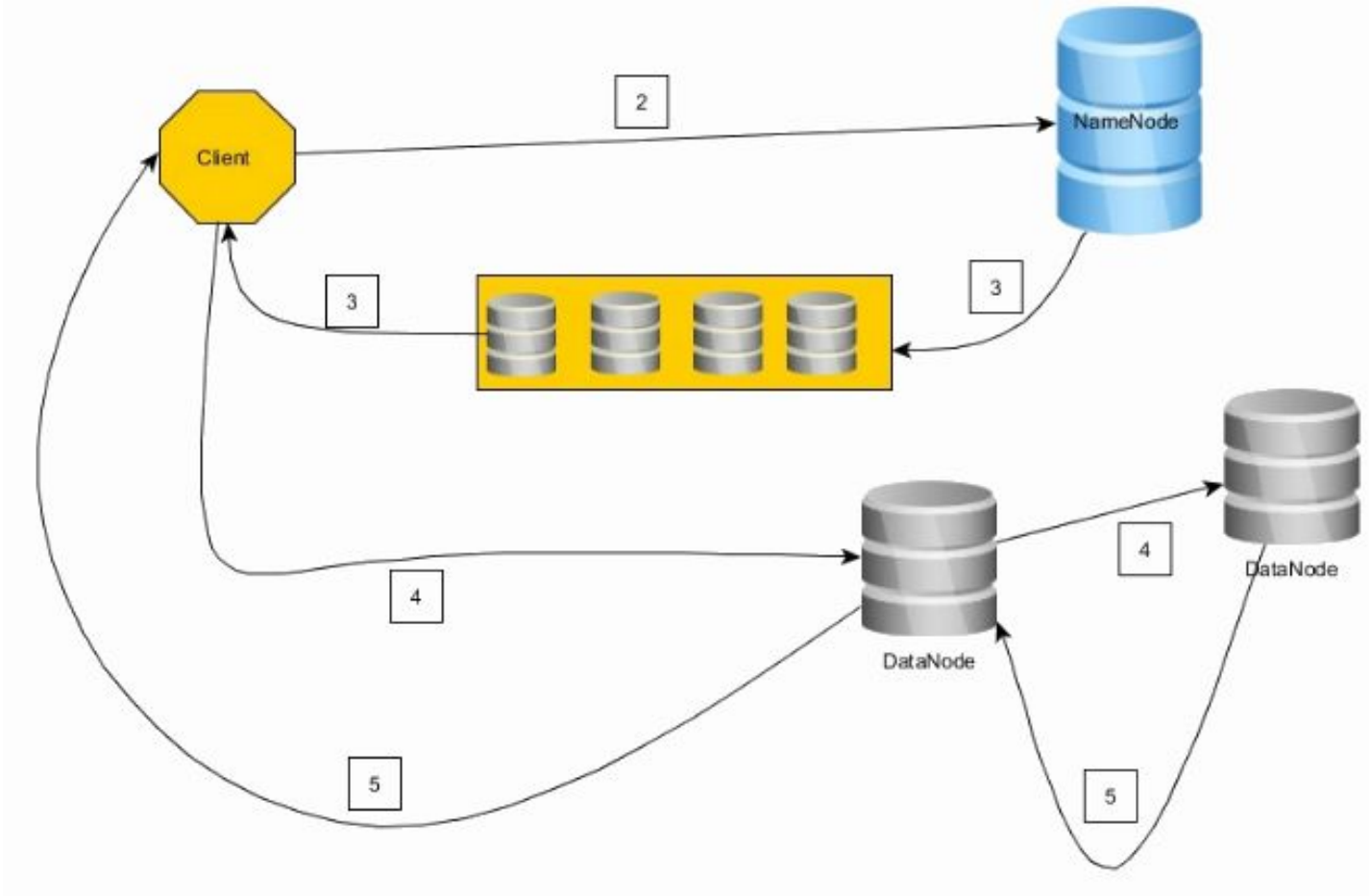
Инкапсуляция сложности реализации

- Hadoop скрывает многие сложности распределенных и многопоточных систем
  - Небольшое число компонент
  - Предоставляет простой и хорошо определенный интерфейс для взаимодействия между компонентами
- Освобождает разработчика от заботы о проблемах системного уровня
  - Race conditions, ожидание данных
  - Организация передачи данных, распределение данных, доставка кода и т.д.
- Позволяет разработчику фокусироваться на разработке приложения и реализации бизнес-логики

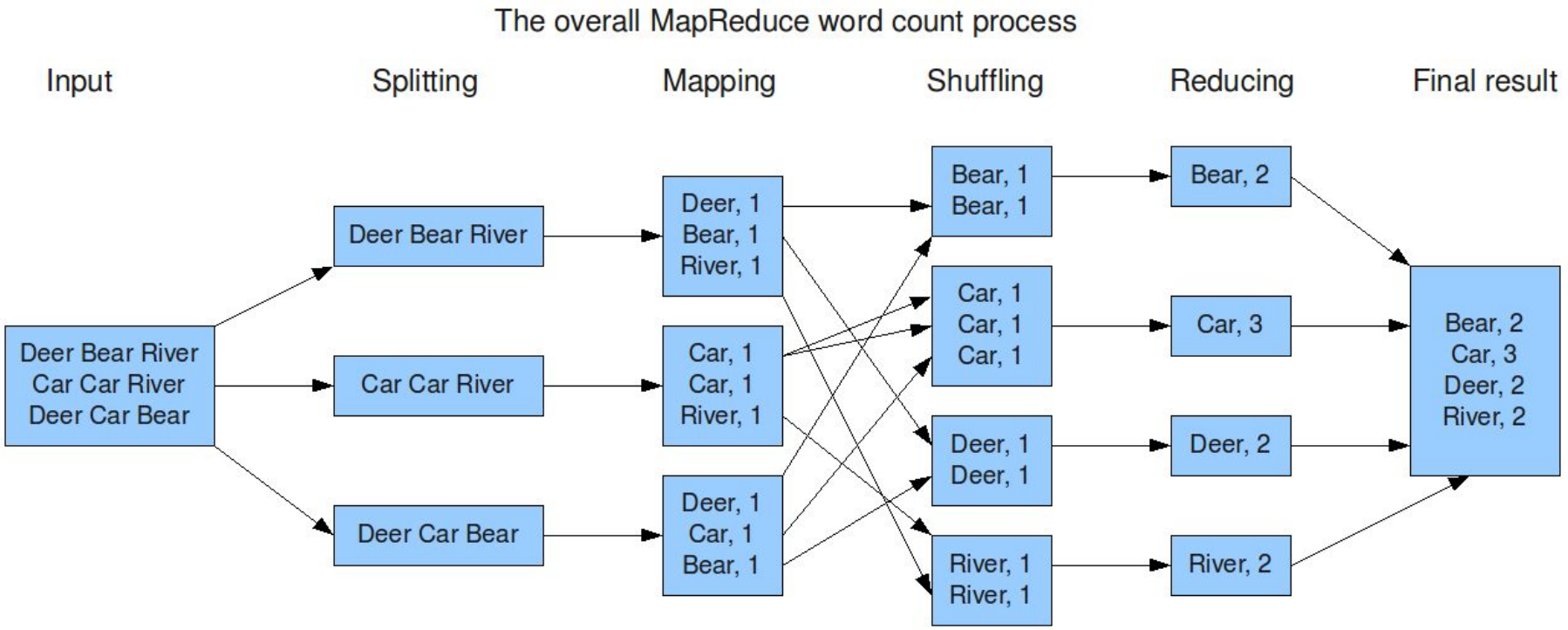
HDFS


master/name
NODE
slave/
data
NODE
slave/
data
NODE
slave/
data
NODE
Master = name NODE
Slave = data NODE

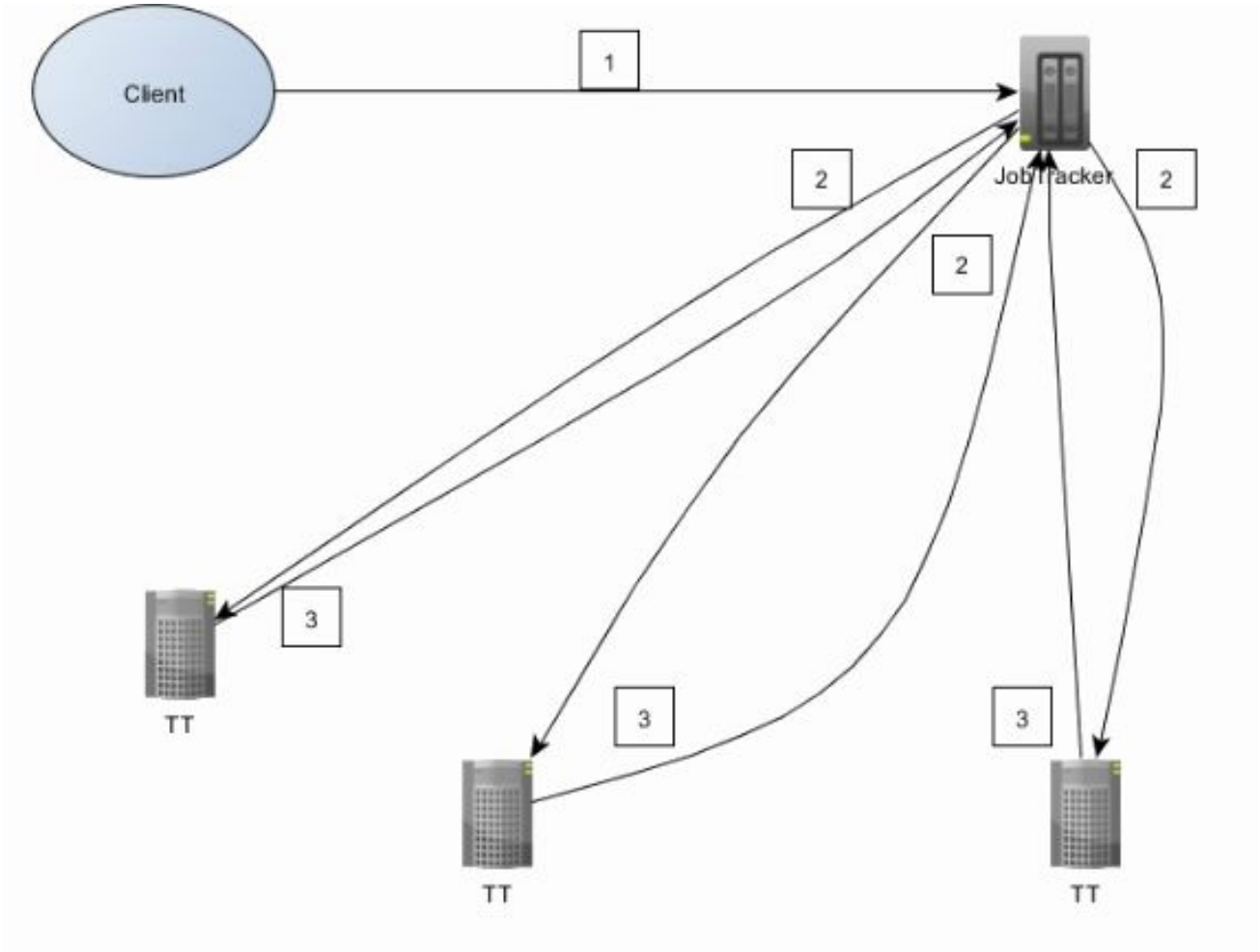
Запись данных в HDFS



Модель MapReduce



Добавление JobTracker и TaskTracker



Внимание!

- Спасибо за внимание!