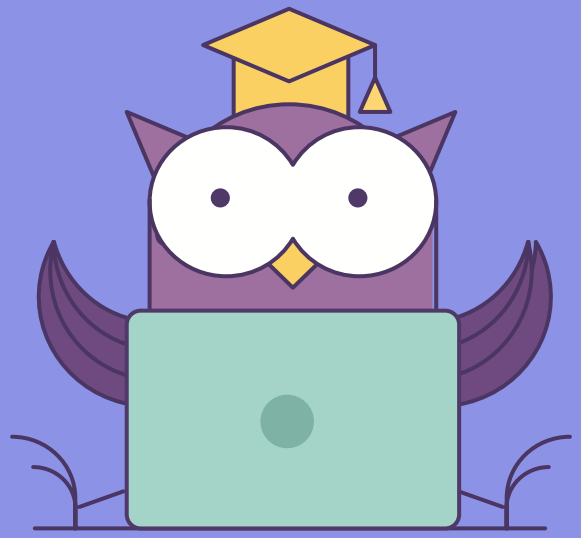
ОНЛАЙН-ОБРАЗОВАНИЕ

Меня хорошо слышно && видно?



Напишите в чат, если есть проблемы!

Ставьте + если все хорошо

OTUS

Дистрибутивы Hadoop

Cloudera и Hortonworks

Иванов Иван Иванович

План на сегодняшнее занятие

- Hadoop и его соседи
- Немного истории
- Cloudera
- Hortonworks

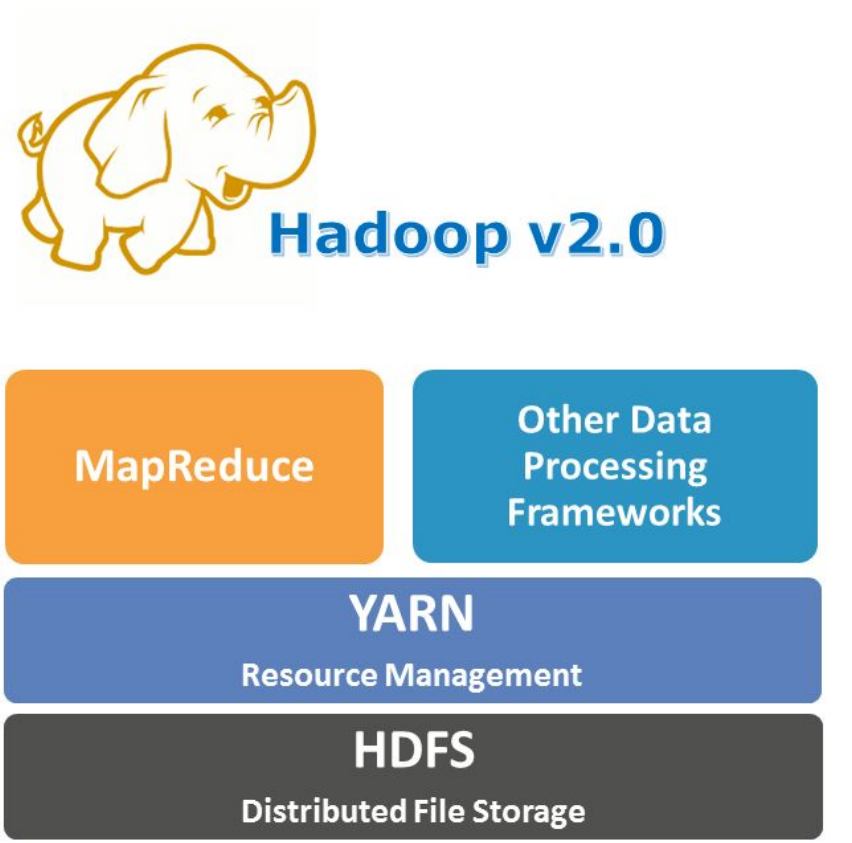
01

Что такое Hadoop

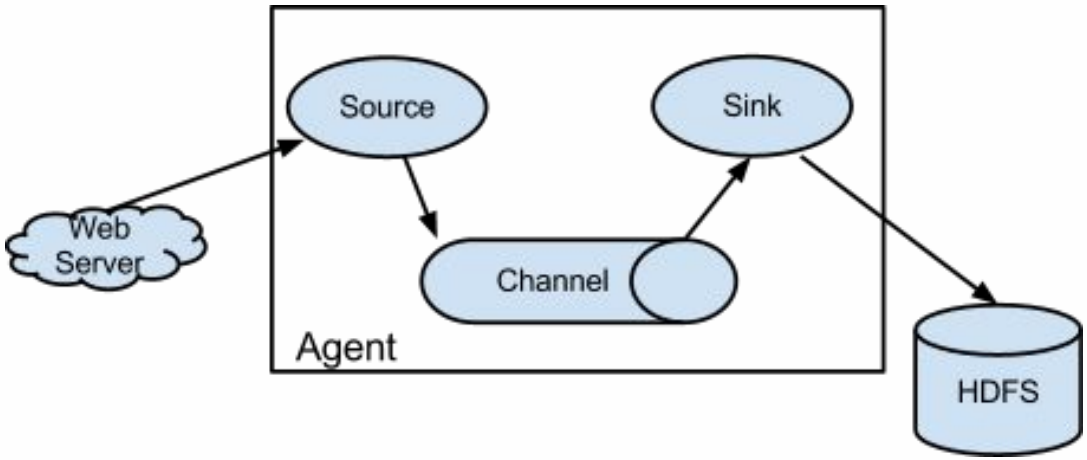
Предпосылки появления Hadoop

- Потребность в распределенных хранилищах
- Масштабирование вычислений
- Управление ресурсами

Классический Hadoop



Apache Flume - инструмент для транспортировки данных между различными источниками и хранилищами.




```
# Name the components on this agent
a1.sources = r1
a1.sinks = k1
a1.channels = c1

# Describe/configure the source
a1.sources.r1.type = netcat
a1.sources.r1.bind = localhost
a1.sources.r1.port = 44444

# Describe the sink
a1.sinks.k1.type = logger

# Use a channel which buffers events in memory
a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000
a1.channels.c1.transactionCapacity = 100

# Bind the source and sink to the channel
a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1
```

Apache Sqoop - по задачам похож на Flume, но ориентирован на загрузку данных из реляционных баз в Hadoop.

Под капотом использует MR.

Sqoop

```
sqoop import --connect jdbc:mysql://localhost/acmedb
--table ORDERS --username test --password ****
```

```
sqoop export --connect jdbc:mysql://localhost/acmedb
--table ORDERS --username test --password ****
--export-dir /user/data/ORDERS
```

Pig

Apache Pig – это высокоуровневый процедурный язык, предназначенный для выполнения запросов к большим слабо структурированным наборам данных с помощью платформ Hadoop и MapReduce. Pig упрощает использование Hadoop, позволяя выполнять **SQL-подобные запросы к распределенным наборам данных**.

Имеет свой язык написания скриптов - Pig Latin.

Пример кода:

```
A = LOAD 'student' AS (name: chararray, age: int, gpa: float);

B = LOAD 'votertab10k' AS (name: chararray, age: int, registration: chararray, donation: float);

C = COGROUP A BY name, B BY name;

D = FOREACH C GENERATE FLATTEN((IsEmpty(A) ? null : A)), FLATTEN((IsEmpty(B) ? null : B));
```

- Позволяет выполнять запросы к слабоструктурированным данным
- Для запросов используется HiveQL
- Имеет свой metastore для хранения “данных о данных” (структур таблиц)

Примеры запросов

```
CREATE TABLE u_data_new (
  userid INT,
  movieid INT,
  rating INT,
  weekday INT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '\t';

SELECT weekday, COUNT(*)
FROM u_data_new
GROUP BY weekday;
```

- Wide-column database
- По сути, KV база данных поверх HDFS
- Хорошо подходит для разреженных данных

Основные понятия

- Table
- Row
- Column family
- Cell

User table Column family for book ratings by userid for bookids

Key	**data**:fname	…	**rating:bookid1**	rating:bookid2
userid1			5	4

Book table Column family for ratings for bookid by userid

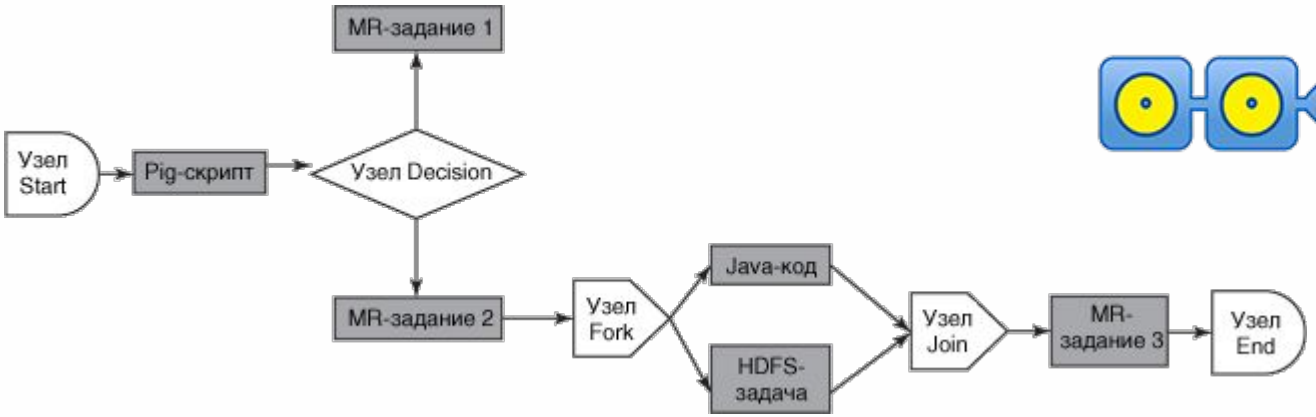
Key	**data:title**	…	**rating:userid1**	rating:userid2
bookid1			5	4

Oozie - родной оркестратор Hadoop

Позволяет

- планировать задачи
- строить сложные пайплайны обработки данных

Интегрирован с MR, Spark, HDFS, Sqoop



02

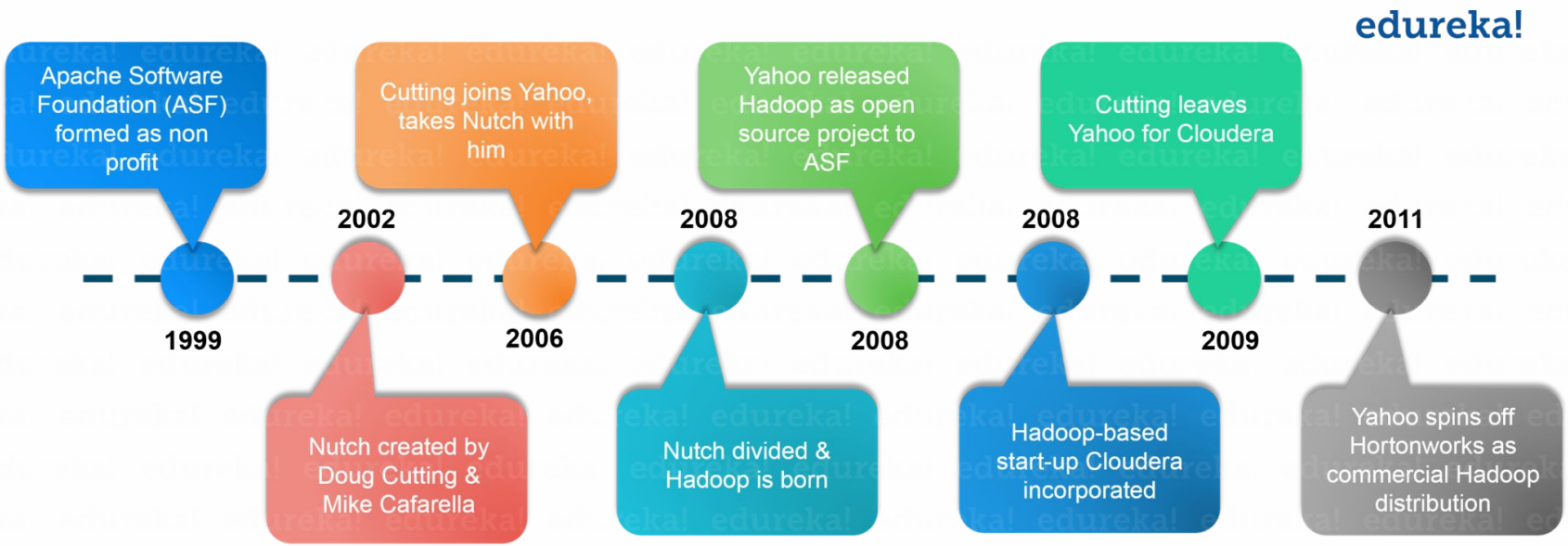
Немного истории

История Hadoop

- 2003 - Google File System paper
- 2004 - MapReduce paper
- 2006 - Hadoop project
- 2008 - Pig, Hive, Zookeeper, HBase
- 2009 - Amazon EMR

Появление “спецов по Hadoop”

- 2008 - создание Cloudera
- 2011 - выделение Hortonworks



Зачем нужны дистрибутивы Hadoop?

- Интеграция компонентов
- Синхронизация версий
- Мониторинг
- Дополнительные инструменты

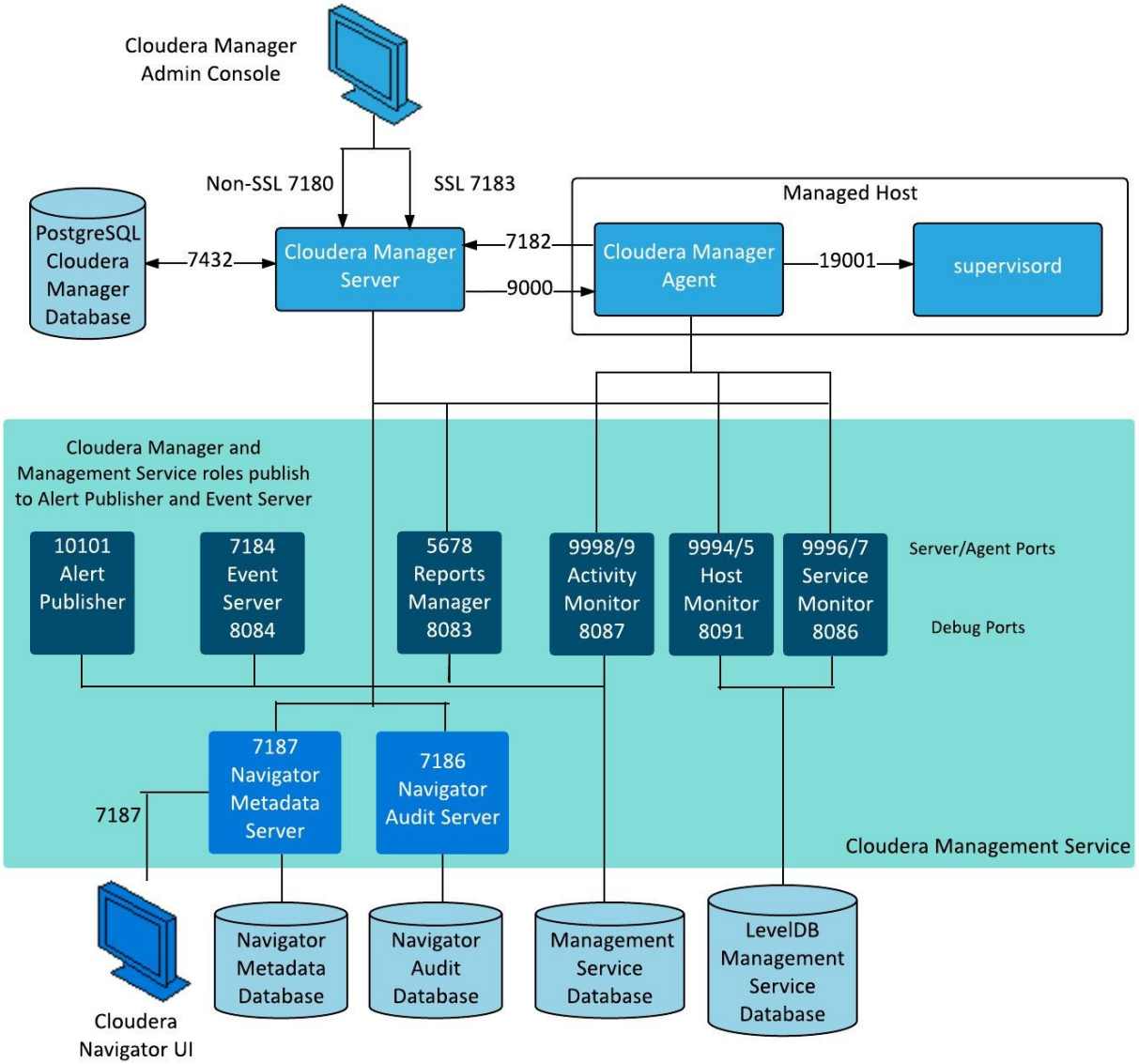
03

Cloudera

Что за продукт у Cloudera?

- CDH - Cloudera’s Distribution including Apache Hadoop (free & enterprise)
- Cloudera Manager
- Cloudera Data Science Workbench

Архитектура Cloudera Manager



Свои инструменты

- Kudu - аналитическое хранилище данных (с ориентацией на time series данные)
- Impala - движок для ad-hoc аналитики
- Sentry - безопасность

Развертывание “боевого” кластера CDH

OTUS

1. Подготовить сервера и создать базы данных
2. Установить Cloudera Manager
    a. Добавить репозиторий Cloudera
    b. Установить пакет
3. Через Cloudera Manager
    a. Add cluster
    b. Add hosts (указать машины и предоставить ключ пользователя с sudo-доступом)
4. Добавить на новые машины роли сервисов

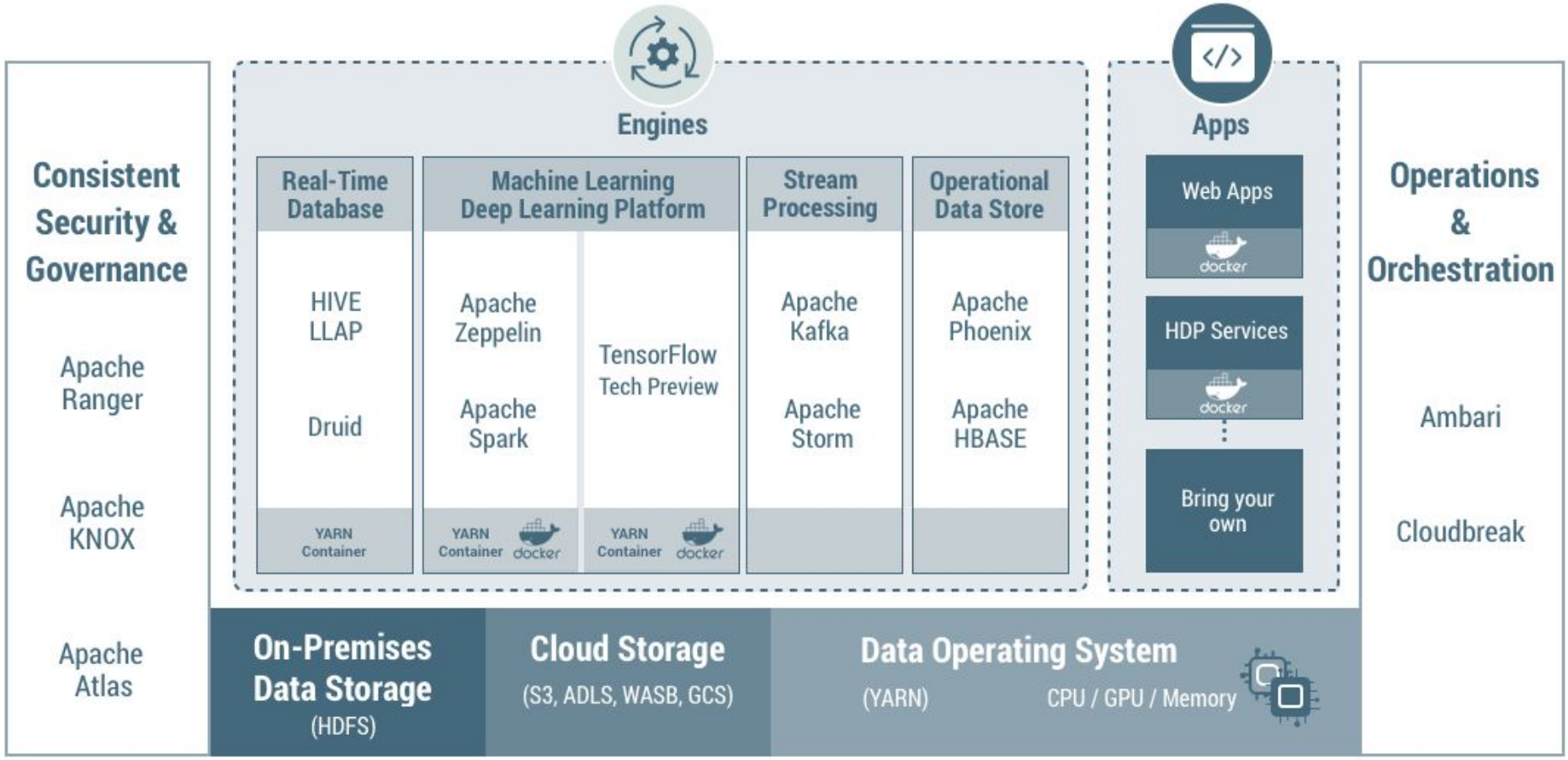
04

Hortonworks

Что за продукт у Hortonworks?

- Hortonworks Data Platform (HDP)
- Hortonworks DataFlow (HDF)
- Ambari (Cluster management)

Что за продукт у Hortonworks?



Свои инструменты

- Knox - proxy для безопасного доступа
- Phoenix - аналитическая база поверх HBase
- Ranger - контроль доступа к Hadoop
- Atlas - data linage (карта данных)
- Ni-Fi - data ingestion tool

05

Практика

Заполните, пожалуйста, опрос в личном кабинете!

Полезные ссылки

- History of Hadoop
- Описание процесса установки CDH

Егор Матешук

egor@mateshuk.com

Спасибо
за внимание!