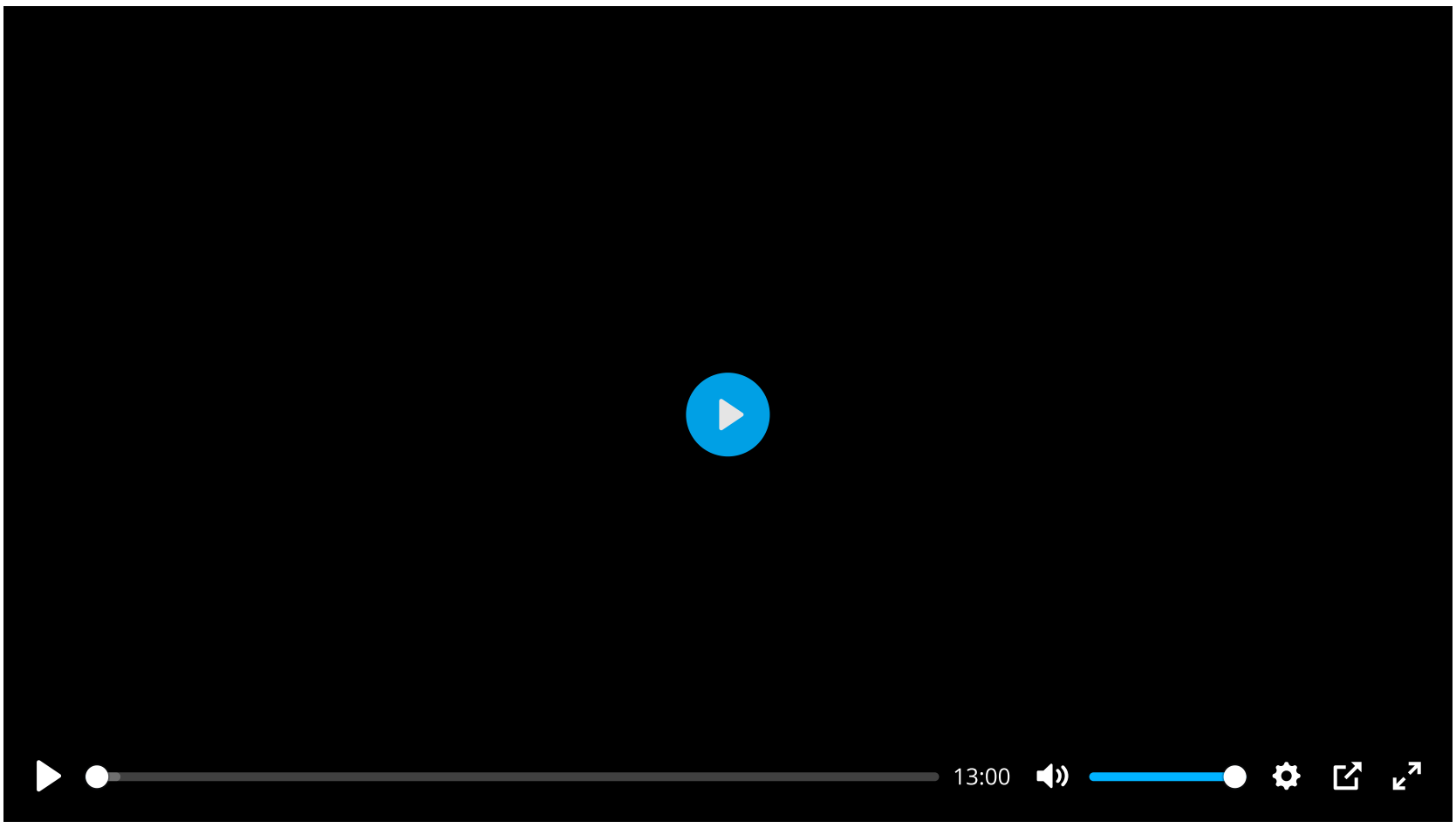
3.2 Экосистема Hadoop – краткий обзор и возможности



Текстовая расшифровка видео:

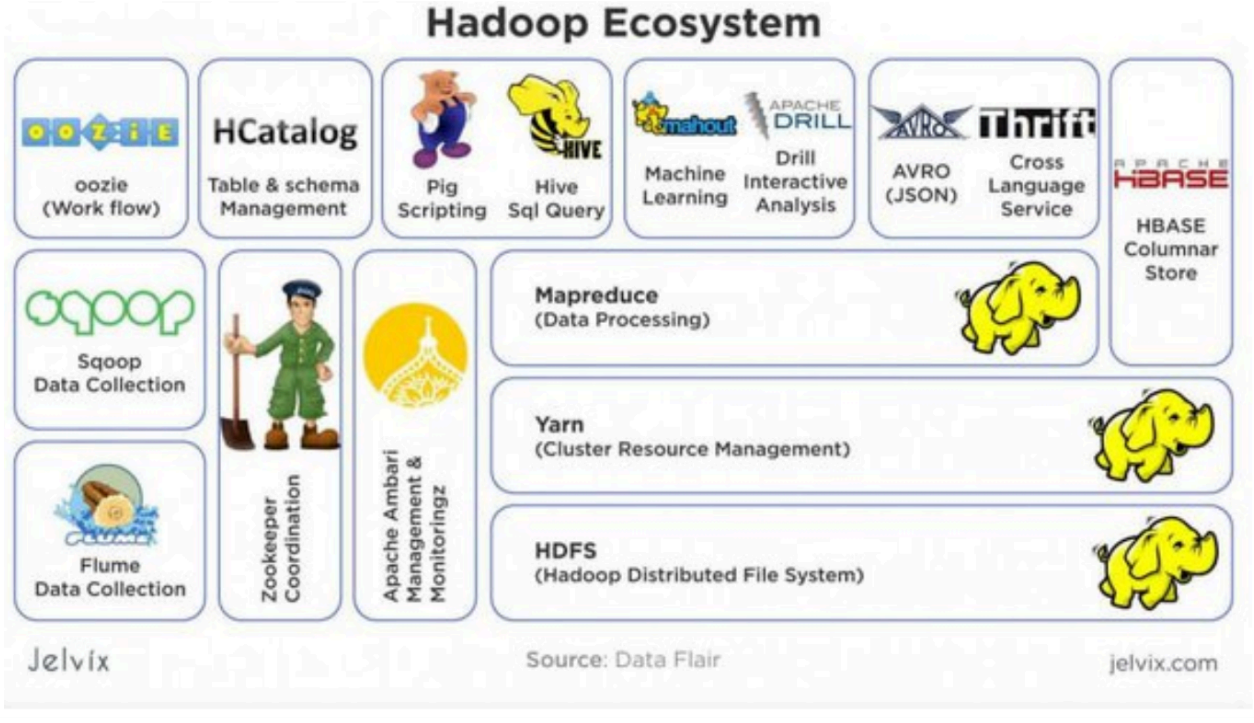
ОСНОВНЫЕ ПРОЕКТЫ ЭКОСИСТЕМЫ HADOOP

План:

- Много составных частей;
- Вендорные решения;
- Тыкаем галочки;
- Надежда Open Source – Apache Ambari;
- Сбор данных по классике;
- Актуальное сегодня;
- Управление зоопарком.

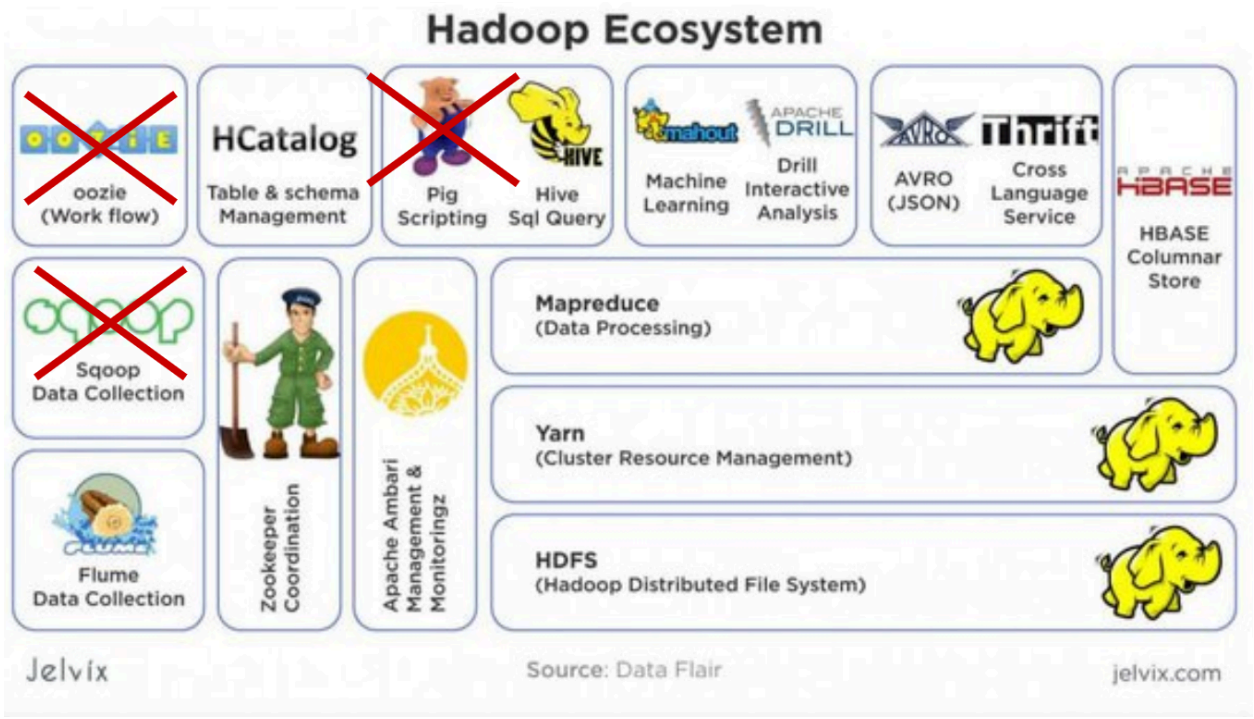
Много составных частей

Если загуглить Hadoop и зайти в раздел изображений, то можно увидеть следующее:



Это большое количество компонентов, отвечающих за разные задачи. В этом легко потеряться.

Часть проектов на сегодняшний день потеряла актуальность:



Вендорные решения

Как получить Hadoop?

Исторически множество движущихся частей, в том числе инфраструктурных, разворачивались при помощи вендорных решений. Поскольку все решения были опенсорсными, их можно было самостоятельно скачать и поставить, однако это было нетривиально.

Некоторые компании представили вендорный Hadoop:

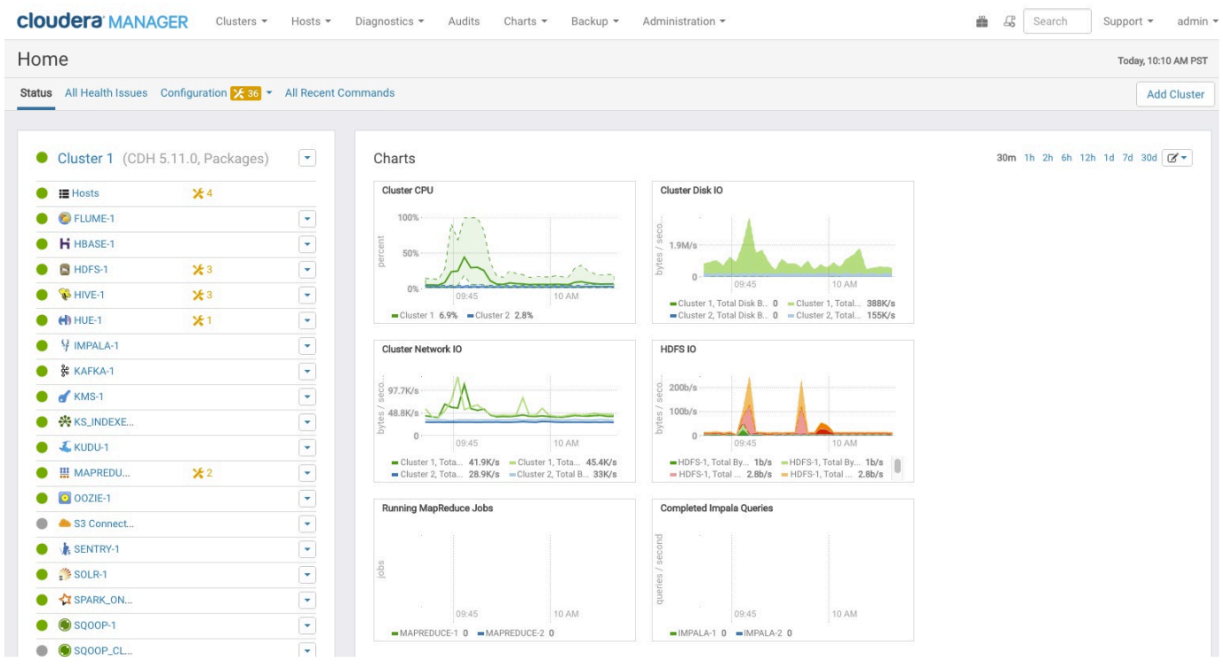
- Cloudera (фактически единственный представитель вендорного Hadoop);
- Hortonworks (был куплен Cloudera);
- Mapr (отошел от дел).

Не стоит забывать об облачных решениях:

- Amazon EMR;
- Azure HDinsight.

Тыкаем галочки

Вендорные решения могут продолжить условный дашборд:

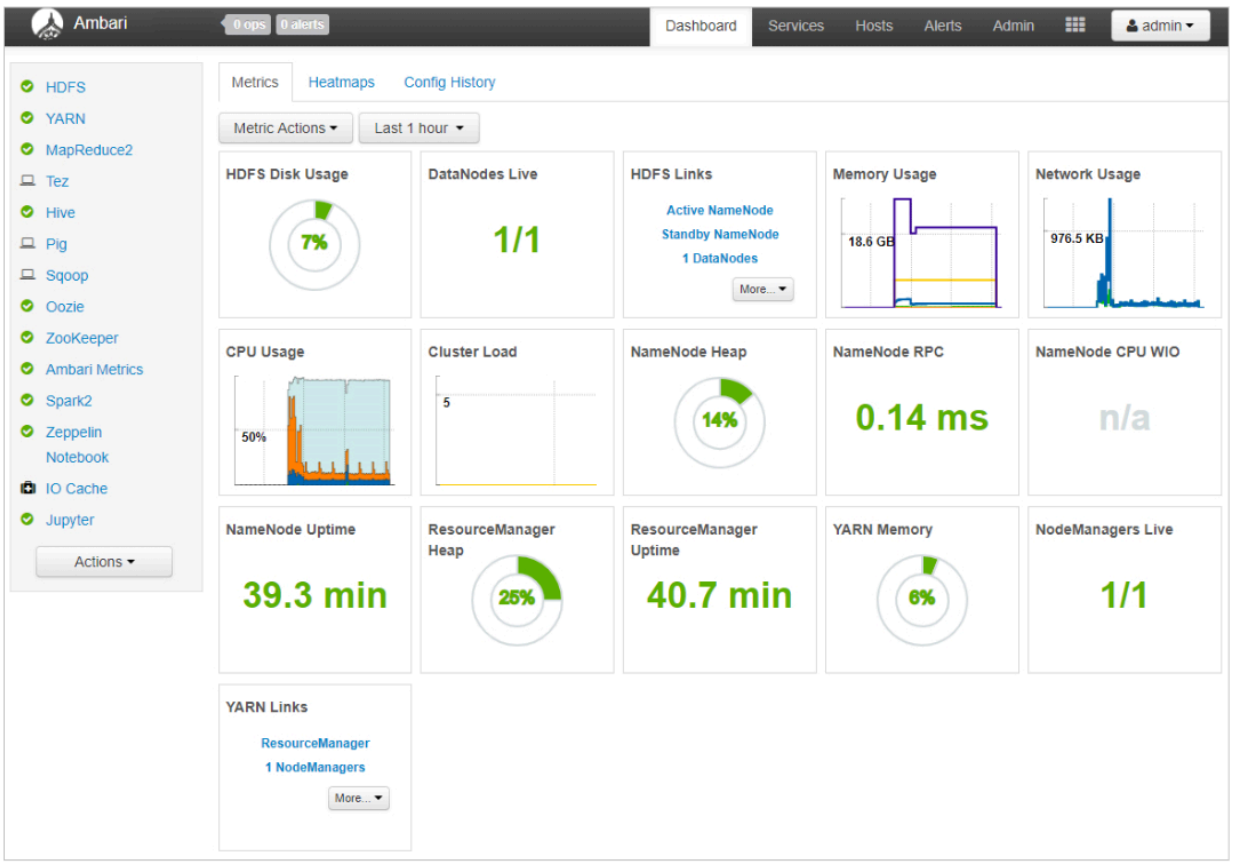


Основная идея: мониторинг состояния кластера.

Надежда Open Source – Apache Ambari

Существует опенсорсная версия дашборда для развертывания мониторинга – **Apache Ambari**.

Пример дашборда:



Предложенный вариант наиболее близок к вендорному.

Сбор данных по классике

Итак, мы развернули Hadoop и поставили компоненты. Что дальше?

Нам нужно достать данные из источников и добавить в Hadoop.

Скорее всего, первое, на что вы наткнетесь, когда попробуете найти способ изъятия данных из продуктовой базы данных, чтобы поработать с ними в аналитическом контуре, будет проект экосистемы Hadoop – **Sqoop**.

Sqoop – редкий пример проекта, в котором новую версию официально упразднили в пользу старой. SQL-запросами он забирает данные из источника и заливает в HDFS.

На сегодняшний день часто используют **Apache Spark**.

Актуальное сегодня

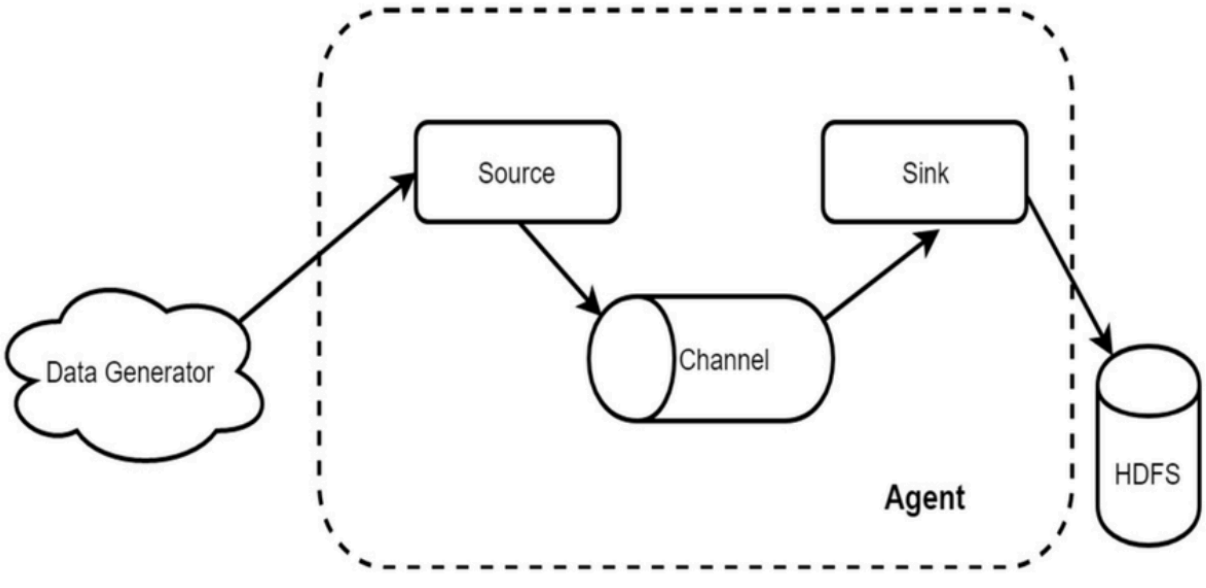
Существует еще один проект, относящийся к экосистеме Hadoop, который называется Apache Flume.

Плюсы Apache Flume:

- Поддерживает стриминг в реальном времени;
- Имеет большое количество коннекторов на вход и выход;
- До сих пор активно развивается.

Архитектура Apache Flume

Рассмотрим схему работы Apache Flume:



Мы можем видеть **источник данных, драйвер** для подключения к источнику (Source). Все это мы пишем в **канал** (это кэш, где хранится последний период). На выходе мы пишем в **драйвер приемника** (Sink) и в **HDFS**.

Альтернативные варианты:

Apache NiFi

Vector (он по умолчанию используется для сбора логов).

Управление зоопарком

Последний компонент, о котором стоит упомянуть, – Apache ZooKeeper.

Apache ZooKeeper – проект, представляющий из себя хитрую базу данных, которая хранит топологию кластера. Благодаря этому проекту получаются хорошие дашборды.