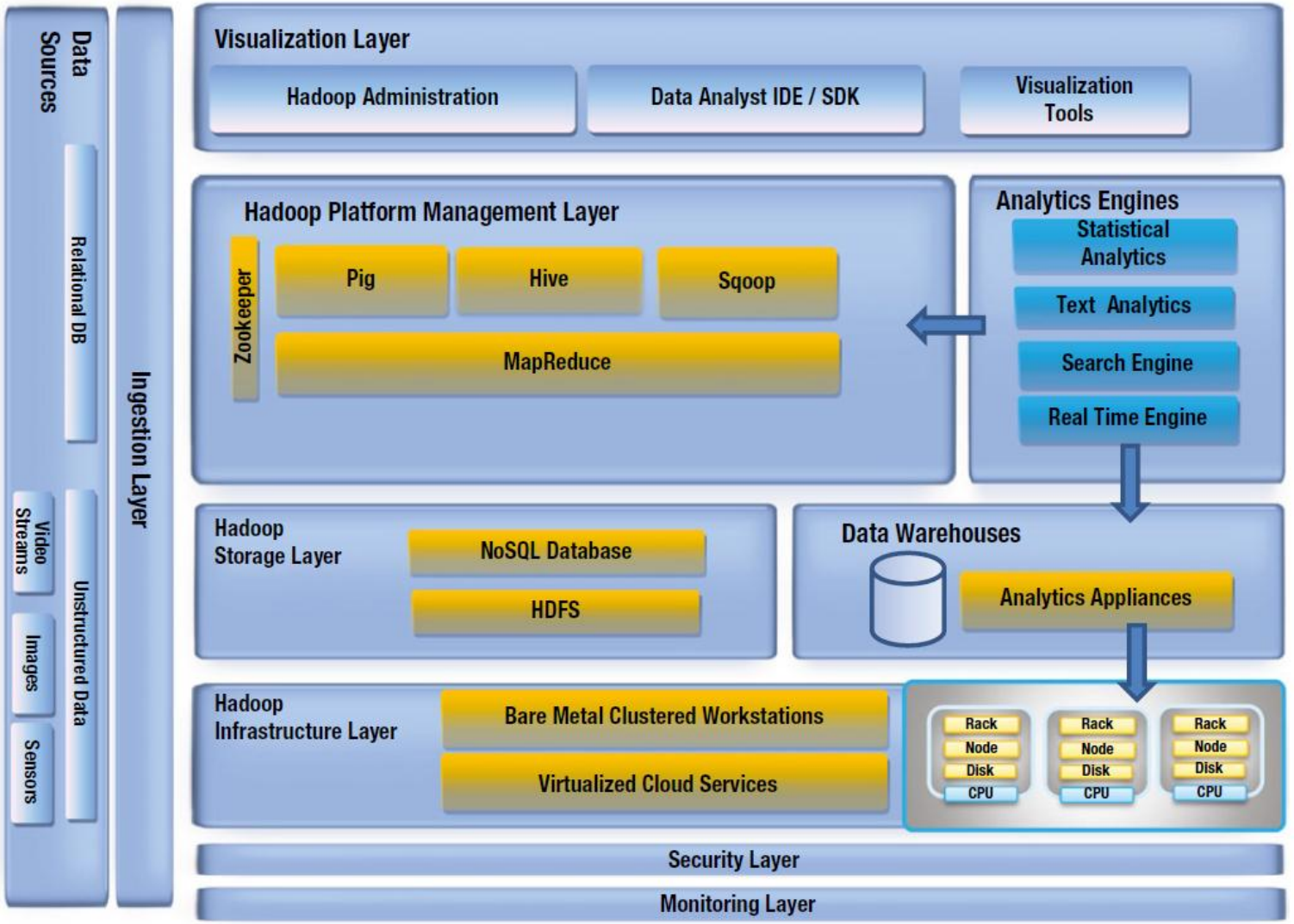
Введение в Big data

Технологии современных СУБД

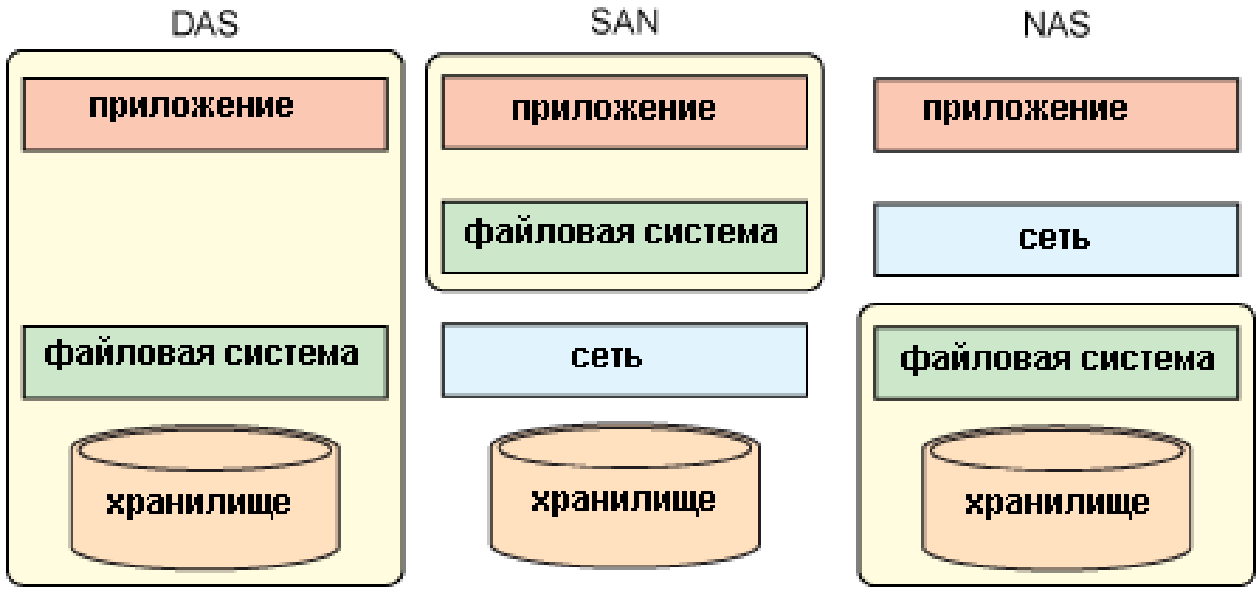
4V

- Volume – объём
- Variety – разнообразие
- Velocity – скорость
- Veracity – достоверность

Архитектура Big data (на примере Hadoop)



СХД



Учесть при проектировании

- Надёжность (доступность, целостность)
- Производительность
- Конфиденциальность

Вопросы, влияющие на выбор технологии

Изменятся ли результаты анализа при изменении набора данных?

- В классификации – вероятно
- При поиске характеристик (получение метаданных о тексте, изображении, к примеру) – маловероятно
- Результат статистических и агрегатных функций – высоковероятно

Допустима ли параллельная обработка?

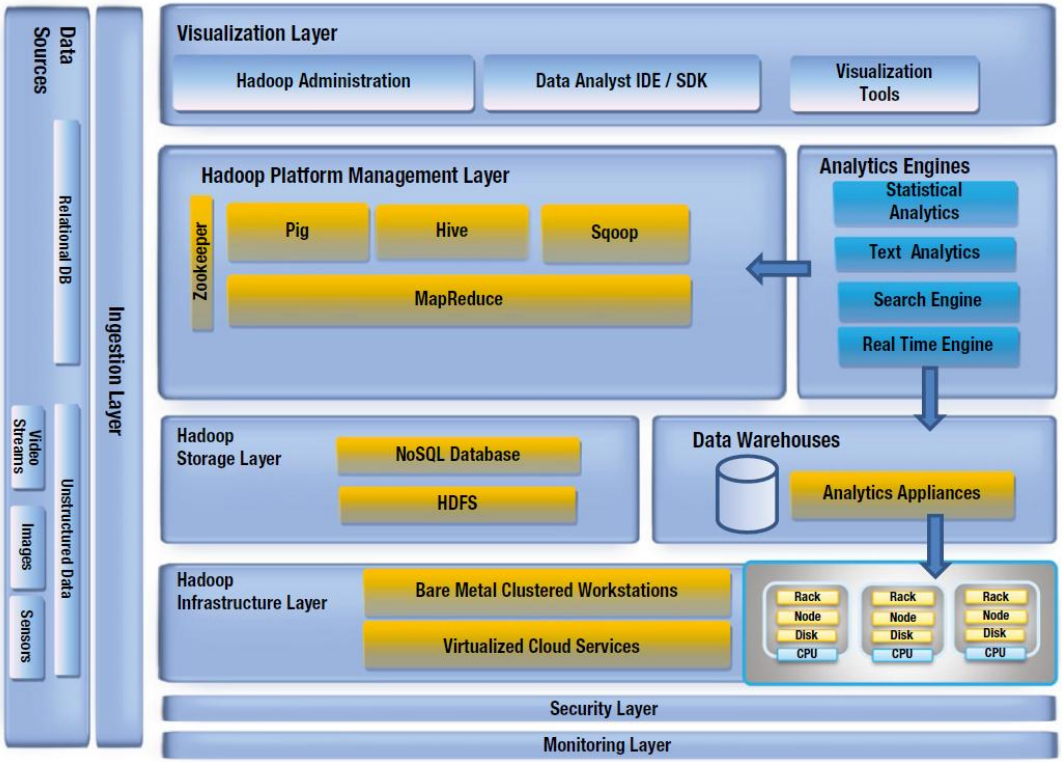
Можно ли для анализа использовать не исходные данные, а их индекс?

Умещаются ли Ваши данные в одно хранилище (HDD\СХД\OpMem)?

HDFS

Файловая система для блочного распределённого хранения файлов большого размера

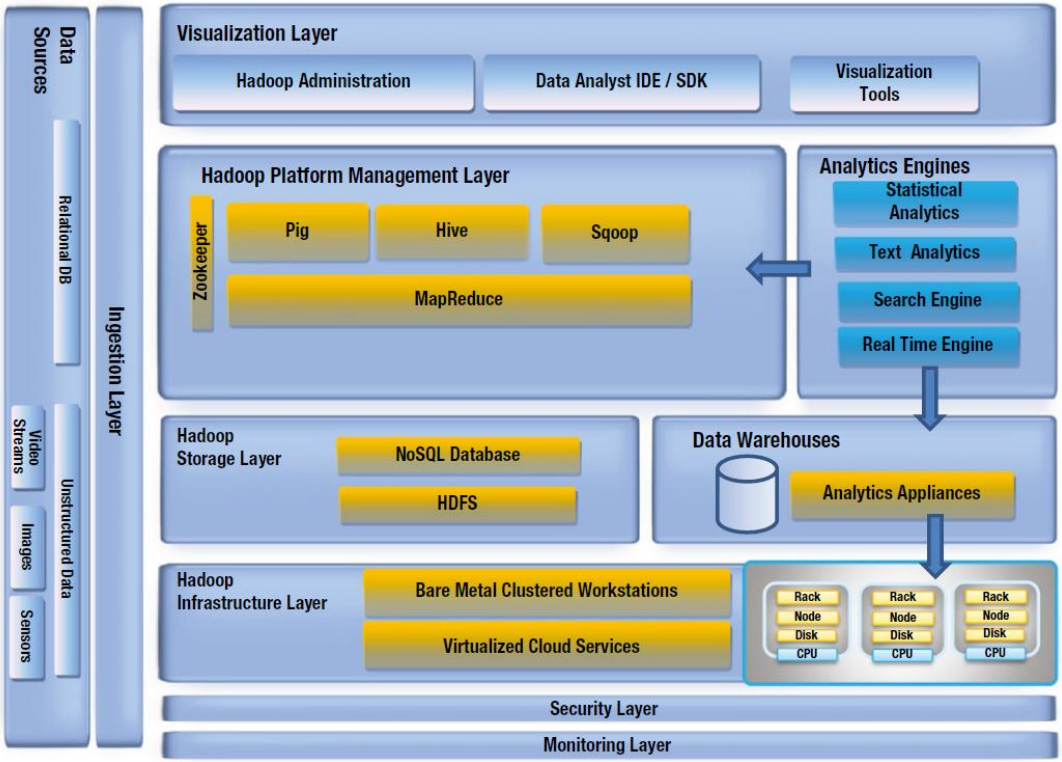
Может использоваться как независимое средство хранения данных



Zookeeper

Средство координации компонент многокомпонентной распределённой системы

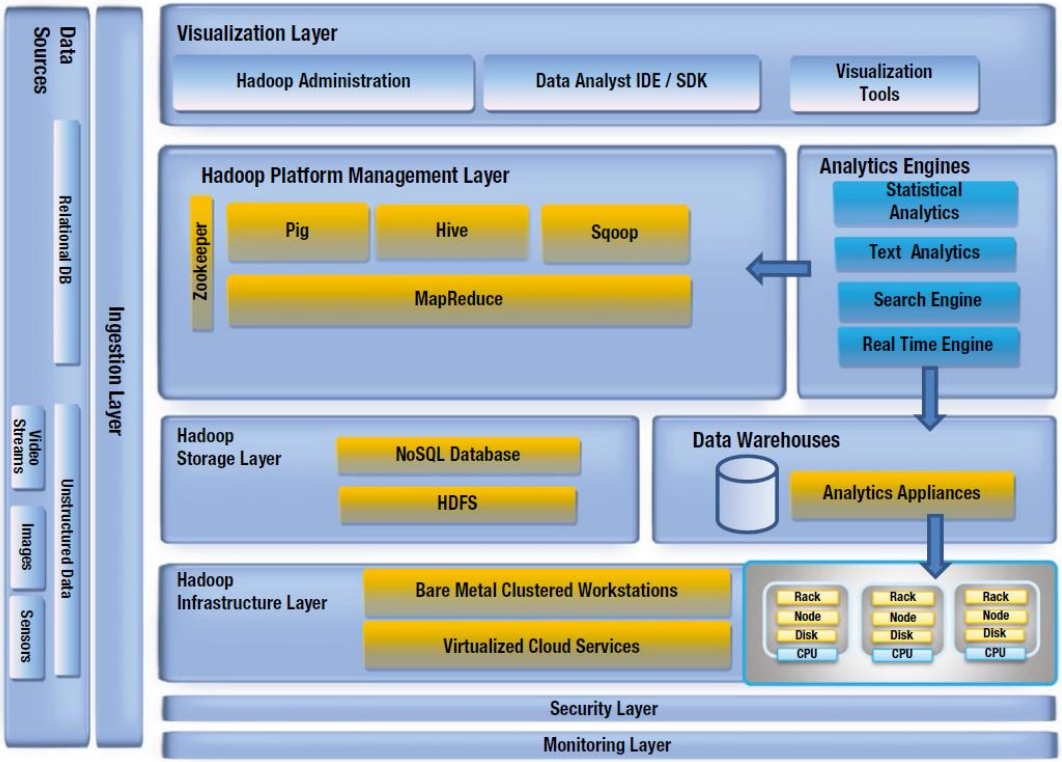
Отвечает за автоматизацию управления жизненными циклами компонент и их синхронизацию



Pig & Pig Latin

Высокоуровневая платформа и язык для манипуляции данными

Абстрагирует от парадигмы распределённых вычислений MapReduce

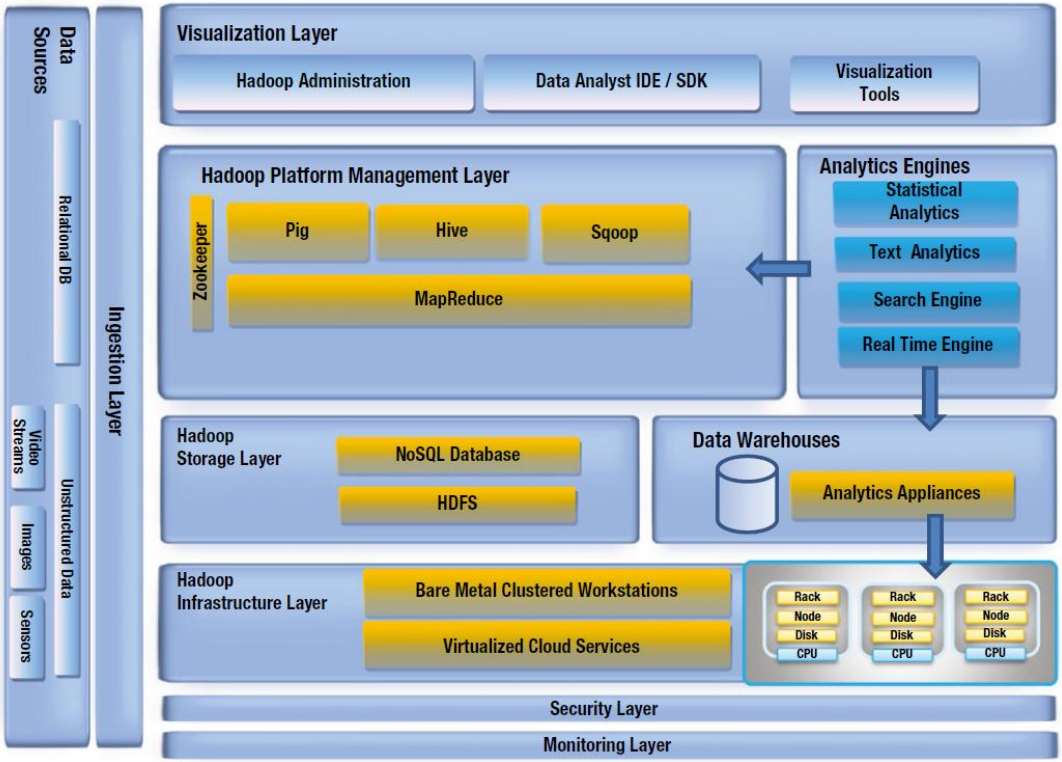


Hive & HiveQL

Инфраструктура хранения и обработки больших объёмов данных

Разработана и внедрена в Facebook, передана сообществу

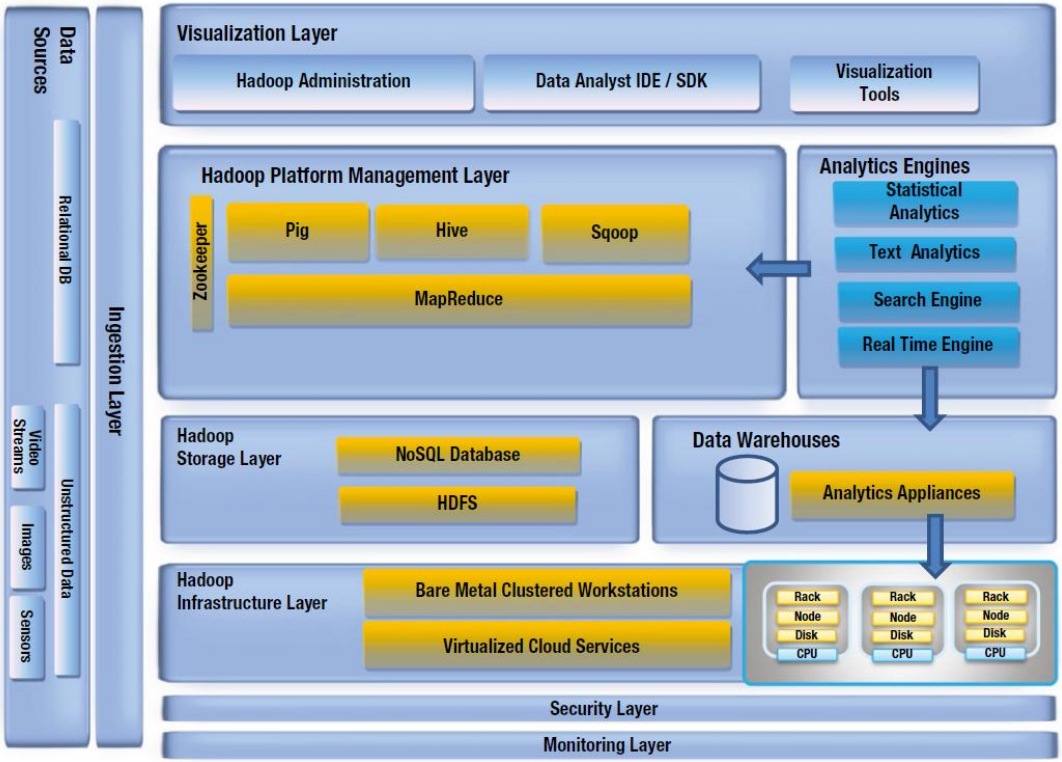
SQL подобный язык управления данных



Sqoop

Утилита командной строки для интеграции и миграции данных между реляционными хранилищами и Hadoop

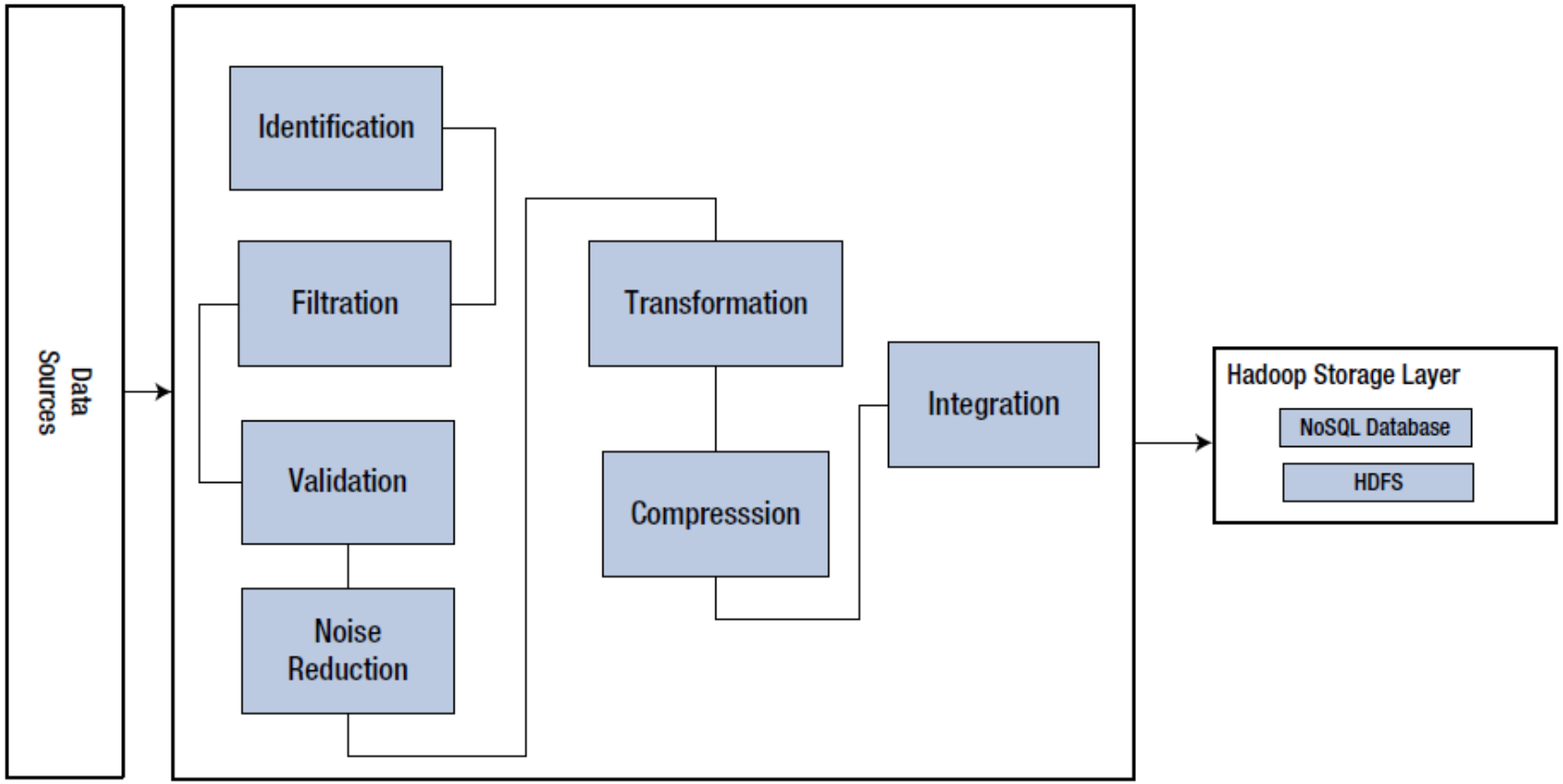
Может использоваться как средство для разработки коннекторов для реляционных СУБД



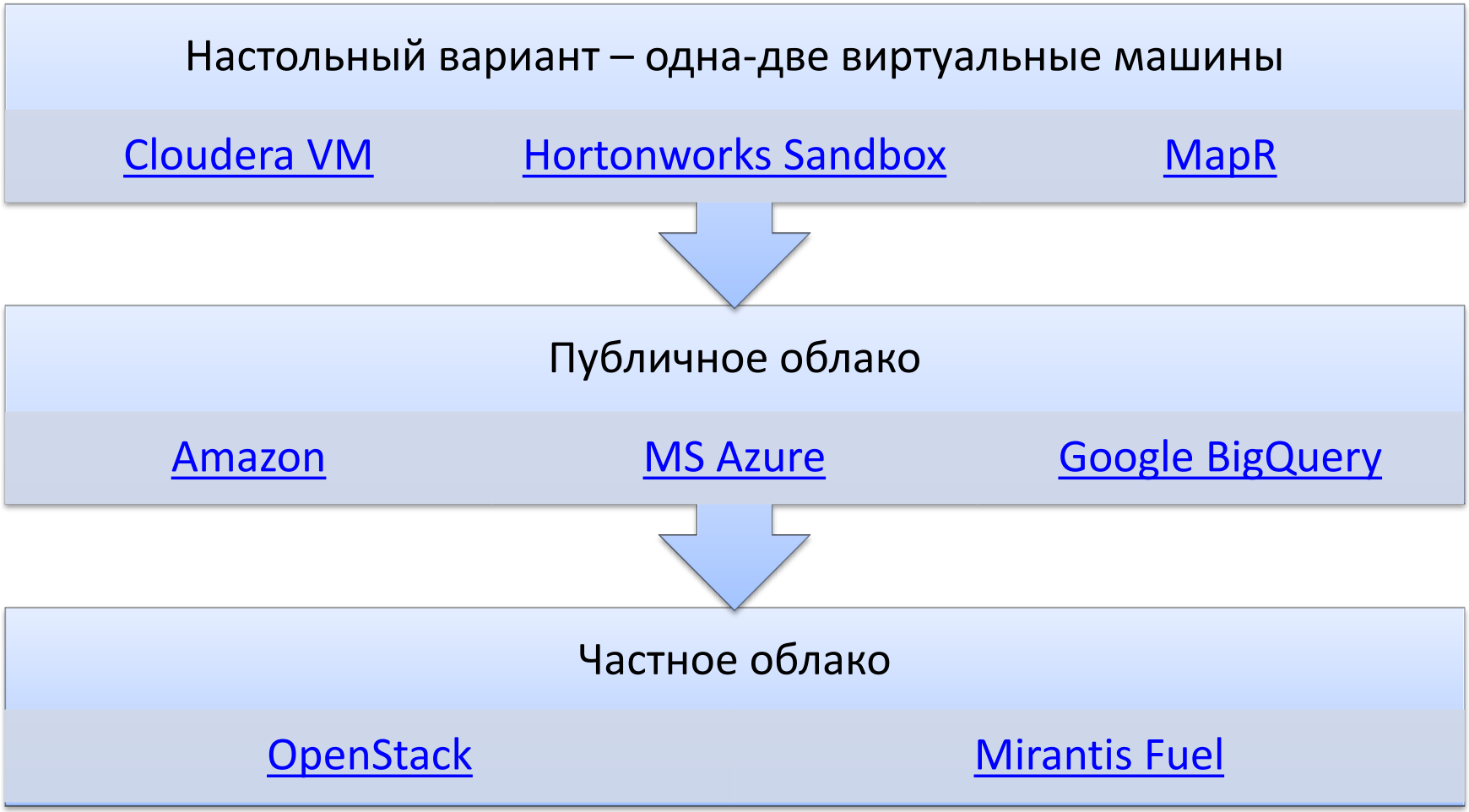
Cloudera

Джеффри Хаммербахер	• Менеджер проекта Hive в компании **Facebook**
Майкл Ольсон	• Вице-президент корпорации **Oracle**, ранее генеральный директор Sleepycat, разрабатывавшей и развивавшей **Berkeley DB**
Амр Авадалла	• Вице-президент корпорации **Yahoo** отвечавший за системы анализа и хранилища данных
Кристофе Бишилья	• Инженер **Google**

Обобщенный процесс Big data анализа



Построение инфраструктуры



Рекомендуемые ресурсы

- Что такое на самом деле Big Data и чем они прекрасны. Лекция Андрея Себранта в Яндексе
- 10 заповедей Больших Данных
- http://bigdatauniversity.com/
- Hadoop: что, где и зачем
- 12 инструментов, о которых необходимо знать каждому программисту, работающему с Big Data (+ см. комментарии к статье)
- Hadoop
  - Hadoop: что, где и зачем
  - Майкл Стоунбрейкер — Hadoop на распутье
  - Утилиты командной строки могут быть в 235-раз быстрее вашего Hadoop кластера
  - Лекции Техносферы
  - Hadoop, часть 1: развертывание кластера
  - Как проиндексировать логи бизнес-приложений в Hadoop (SolrCloud)
- Spark
  - Why Apache Spark is a Crossover Hit for Data Scientists
  - Real-time Data Mining with Spark