Microsoft

DevCon School

Технологии будущего

DevCon School
Технологии будущего

Большие данные на платформе Microsoft Azure

Дмитрий Сошников
Technical Evangelist, Microsoft

Цели

Что такое Big Data

Откуда берутся большие данные и что с этим делать

Обзор технологий

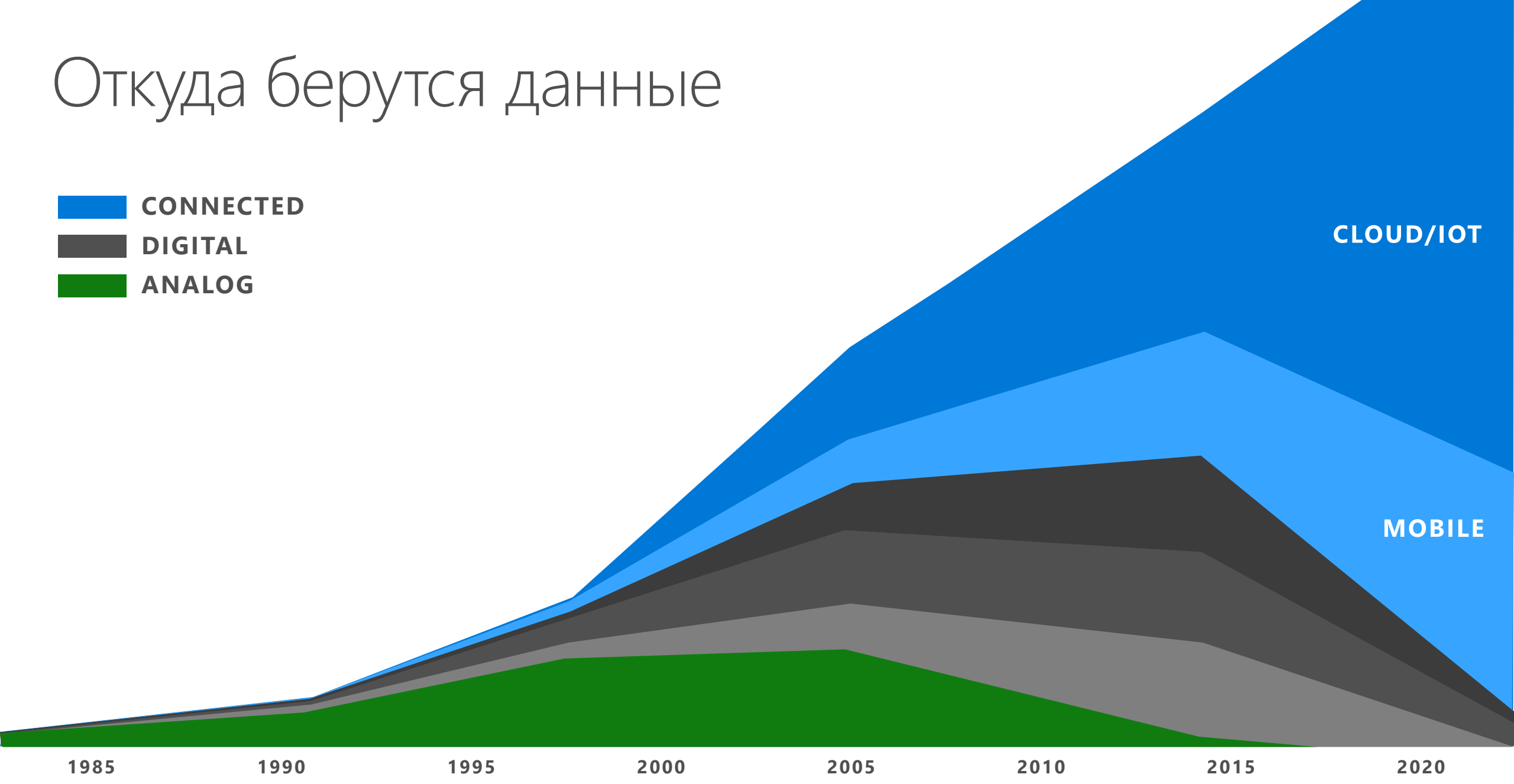
Что есть в Azure для работы с большими данными

Орг.вопросы

Что мы будем делать на школе

Что такое большие данные

Откуда берутся данные



Классификация больших данных

Зачем собирать большие данные

Получение полной картины бизнеса

- Удалённый мониторинг
- Предсказание показателей
- Управление рисками

Распределение продуктов

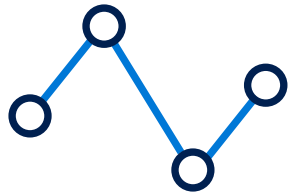
- Inventory management
- Supply chain optimization

Персонализация для конечных пользователей

- Personalized offers
- Product recommendations

Проактивное устранение проблем

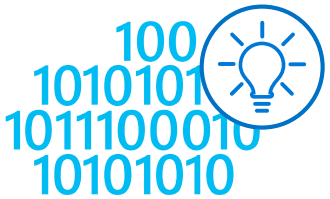
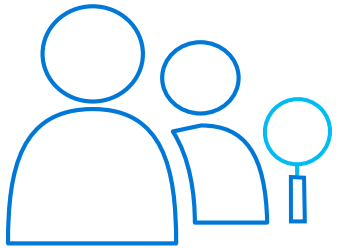
- Predictive maintenance
- Operational efficiency
- Customer service improvement



Новые возможности бизнеса

- Cross-sell and upsell
- Product-as-a-service
- New data-driven services

Современный подход к большим данным

	Традиционный	Big Data
 Структура данных	Реляционные с предопределенной схемой	Слабо-структурированные с гибкой схемой
 Стоимость	Специализированное оборудование	Общего назначения
 Культура	Операционные отчеты Смотрим в прошлое	Пробы и эксперименты Машинное обучение, a/b тестирование и т.д.

Опыт Microsoft

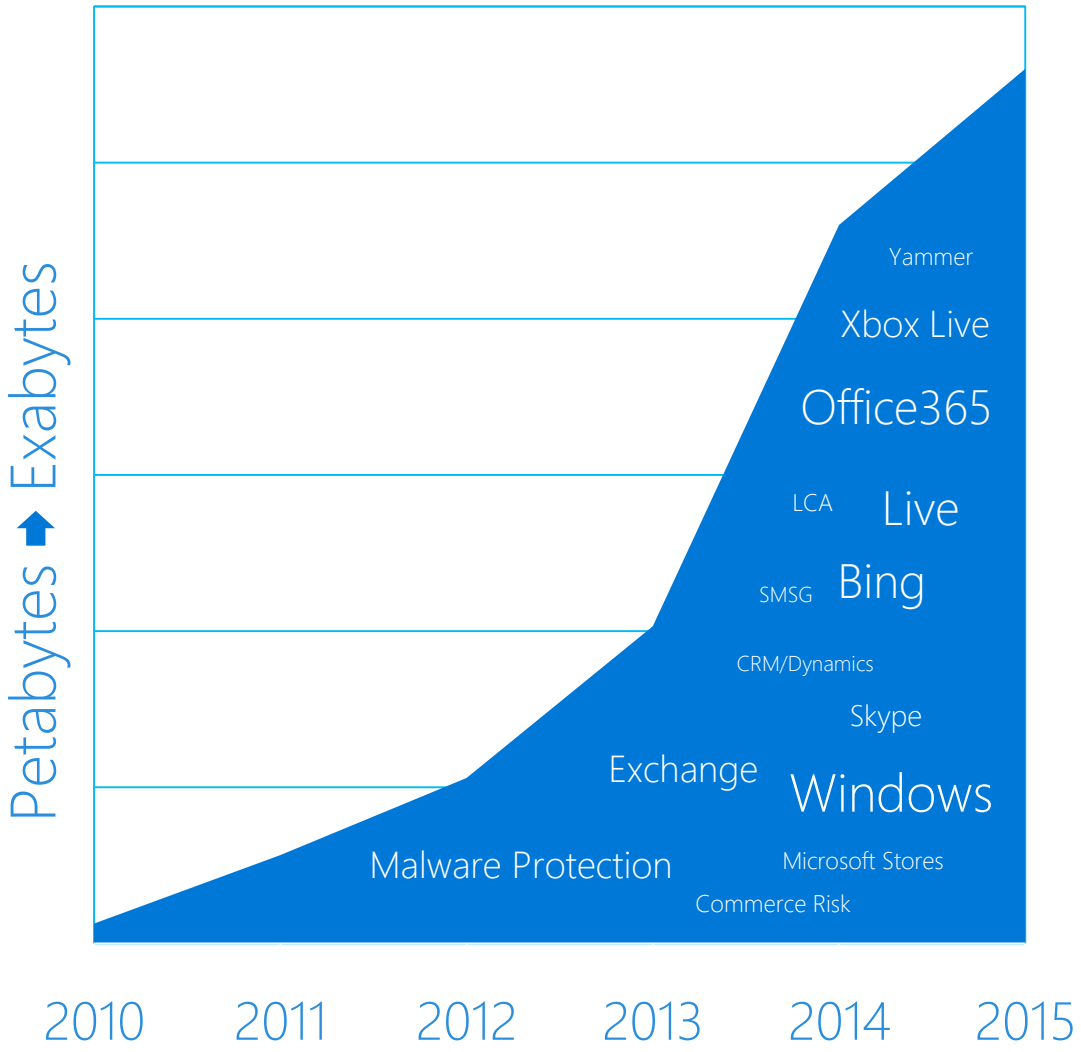
Microsoft столкнулся с большими данными при создании своих продуктов.

Что мы сделали:

- Общедоступные сервисы для больших данных
- Инструменты, доступные разработчикам
- Доступные инструменты машинного обучения

В результате:

- Разработчикам доступны те же инструменты, которые используются внутри Microsoft

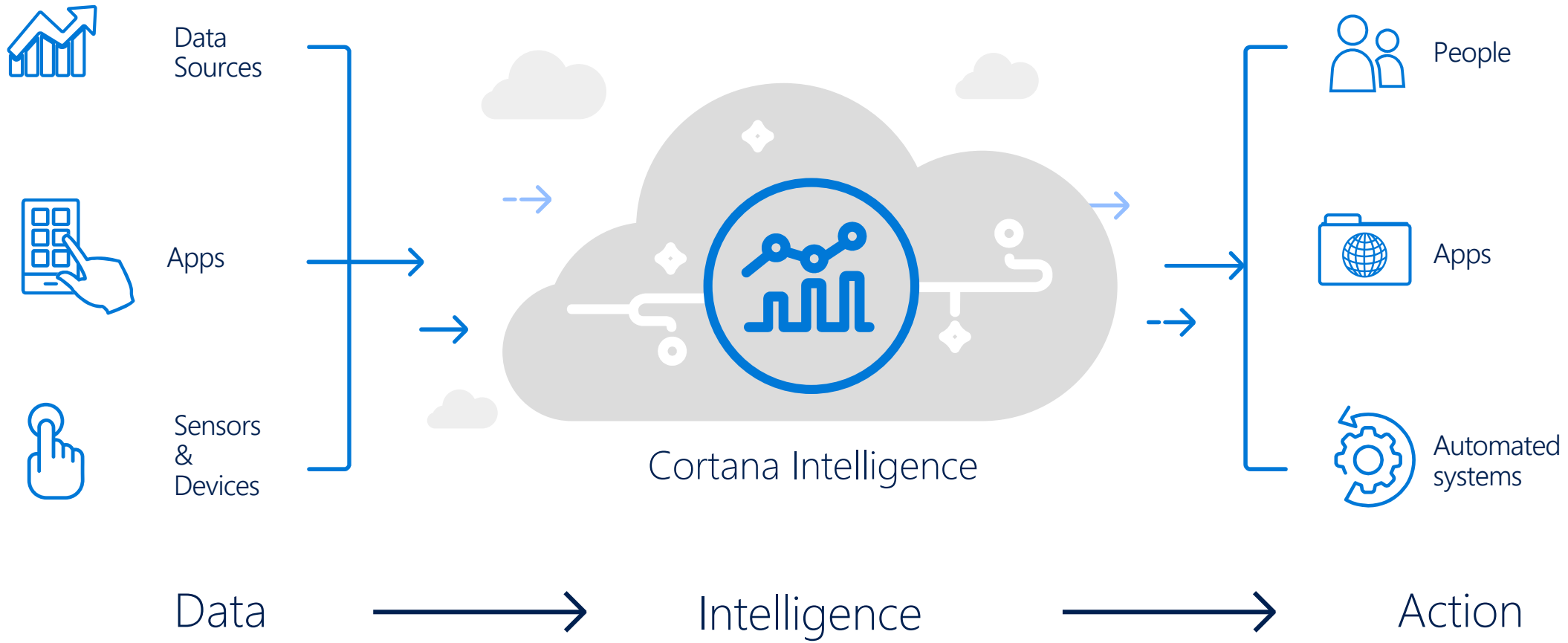


* Microsoft. FY16 Q4 Results, URL: http://www.microsoft.com/en-us/Investor/earnings/FY-2016-Q4/press-release-webcast

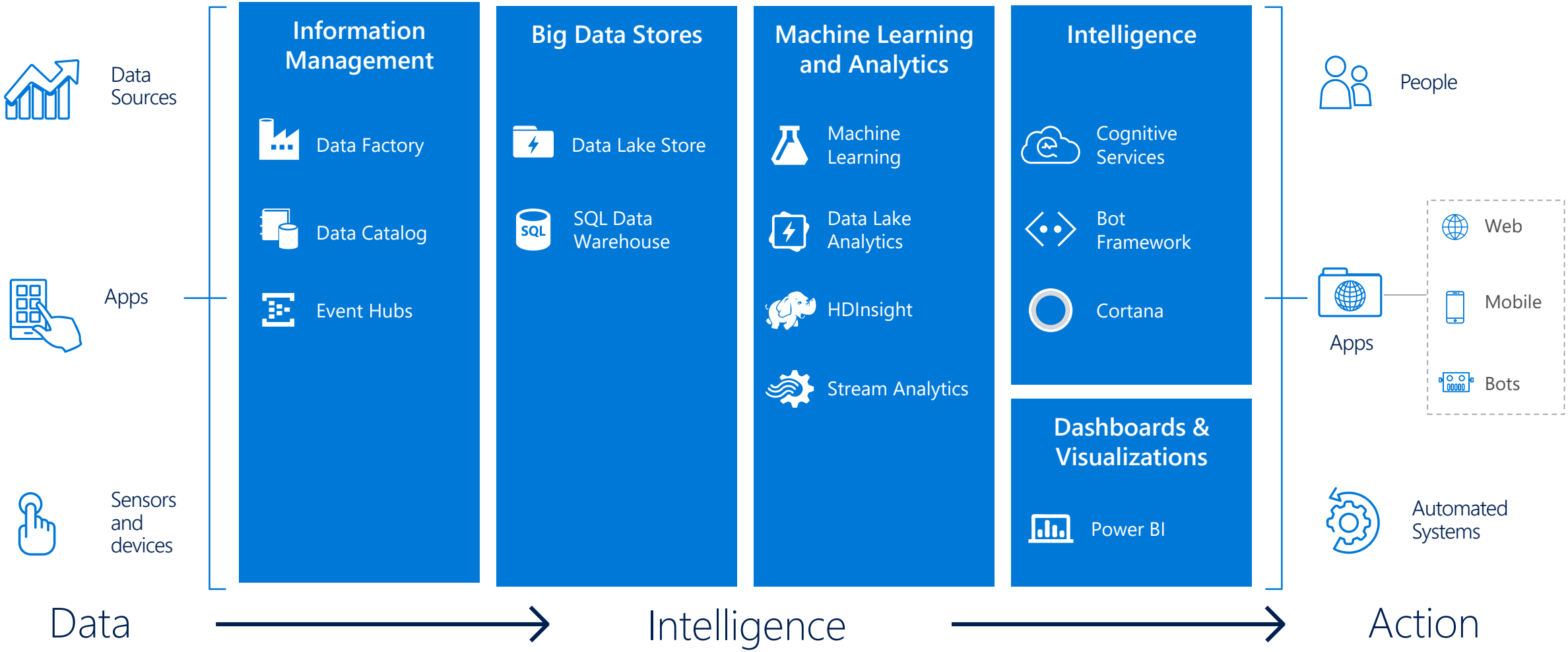
Cortana Intelligence Suite

Cortana Intelligence Suite

Превращает данные в умные действия



Cortana Intelligence Suite



Получение данных

Потоковые данные



Event Hubs

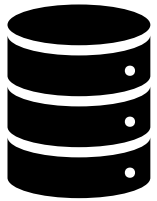
IOT hubs

kafka

Stream Analytics

Azure Function

Перенос данных




Data Factory

Хранилище

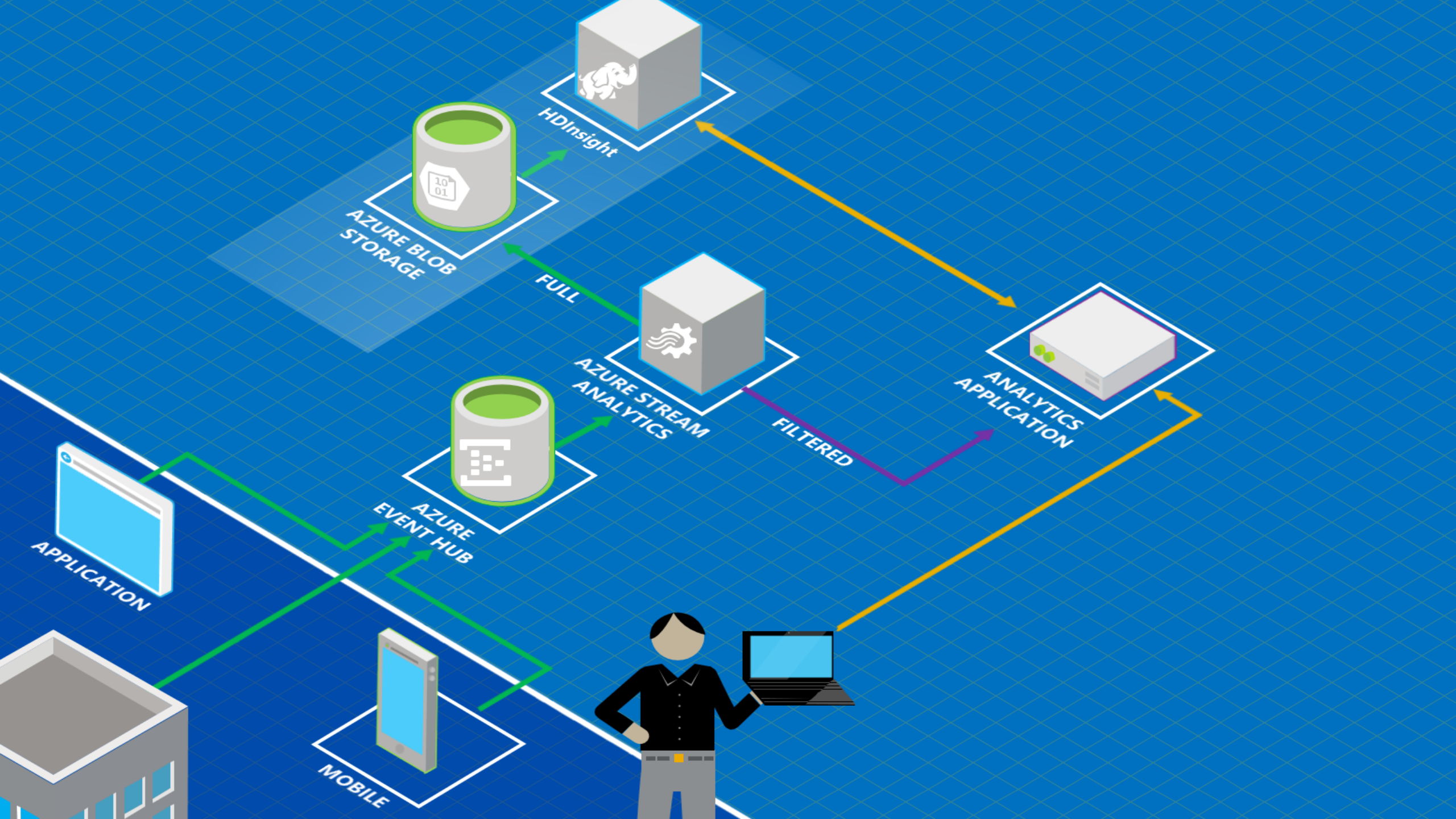
Демонстрация

Собираем IoT-данные

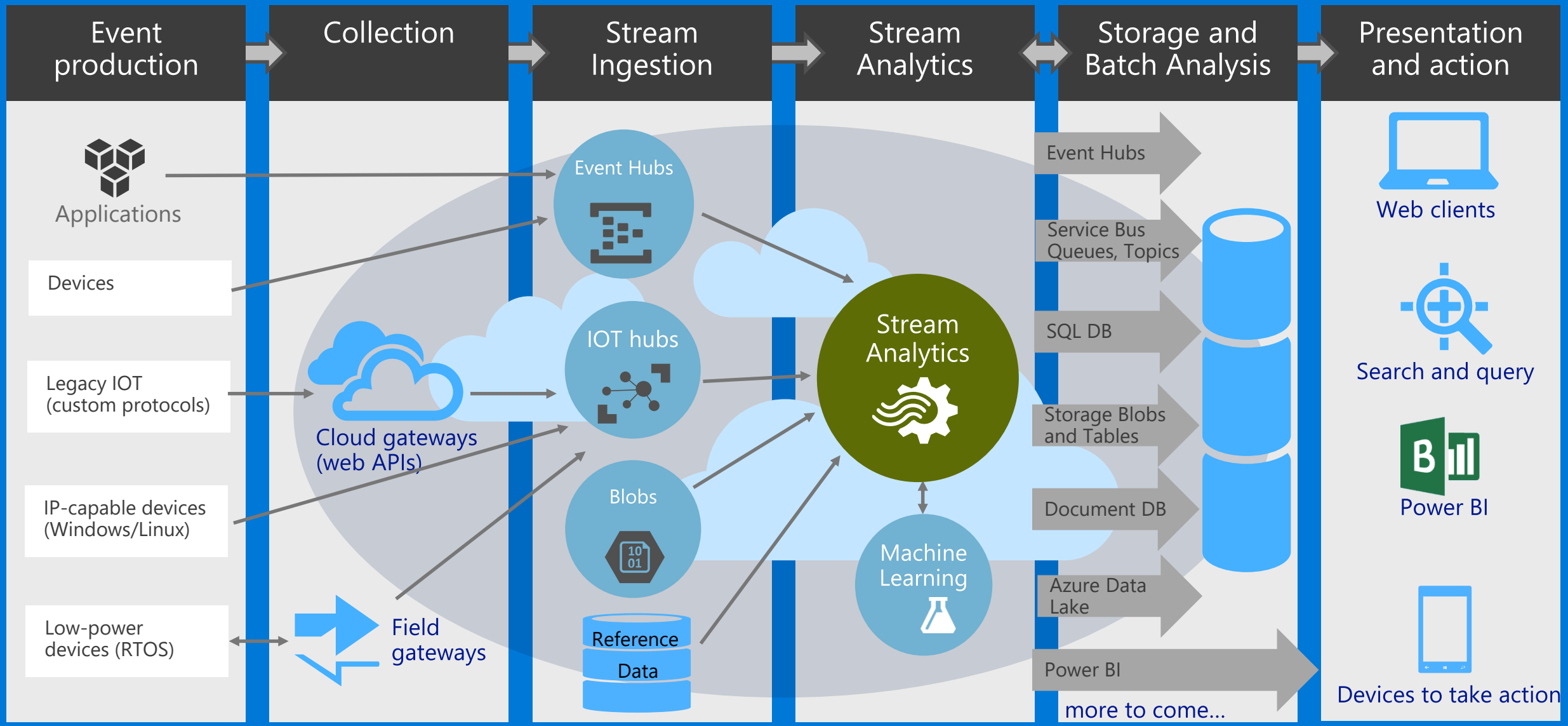
http://github.com/evangelism/FaceRecognitionTracker

#msdevcon

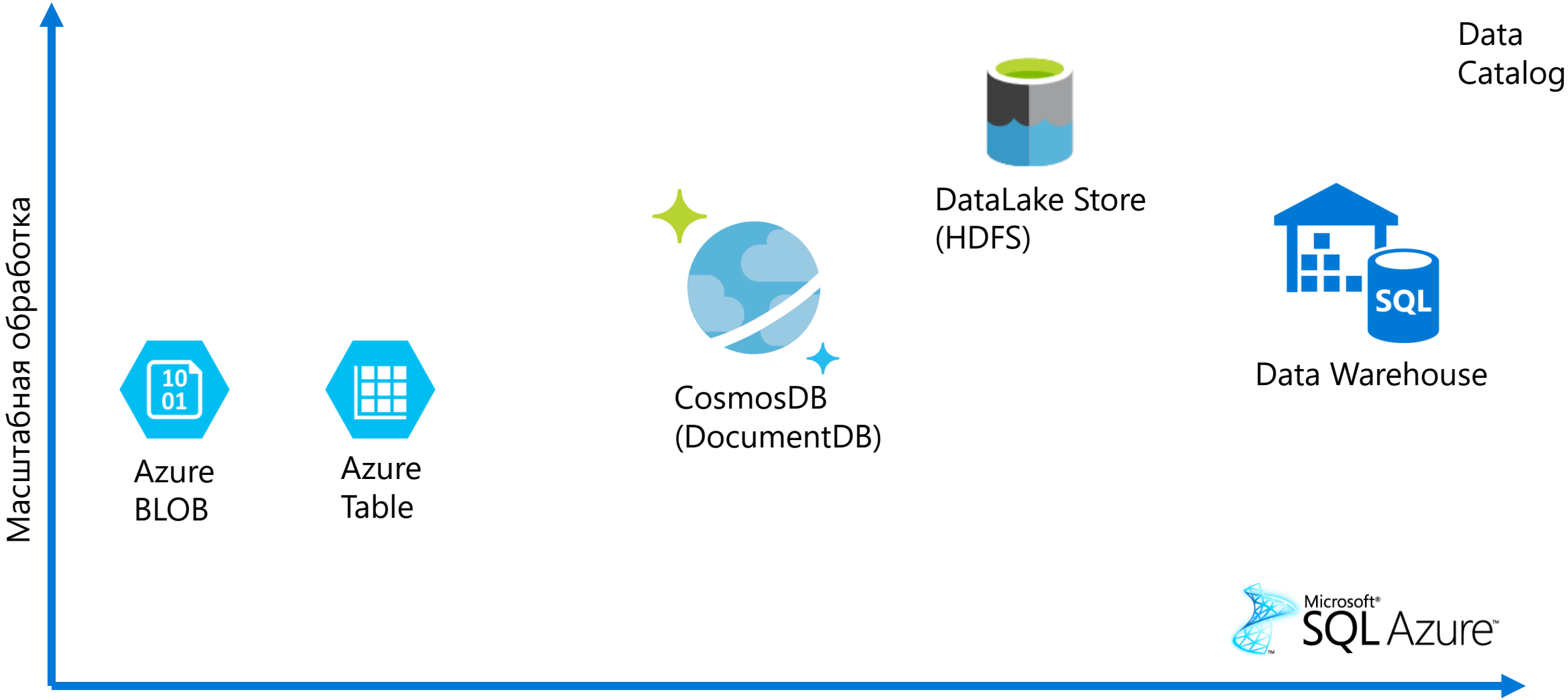
DevCon School
Технологии будущего


HDInsight
AZURE BLOB STORAGE
FULL
AZURE STREAM ANALYTICS
FILTERED
ANALYTICS APPLICATION
AZURE EVENT HUB
APPLICATION
MOBILE

Типичная схема IoT-решения



Хранение данных



SQL

mongoDB

Gremlin

$G = (V, E)$

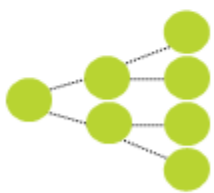
Azure Cosmos DB

... *more coming soon*

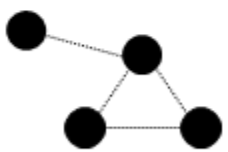
Key-Value

Column-family

Documents



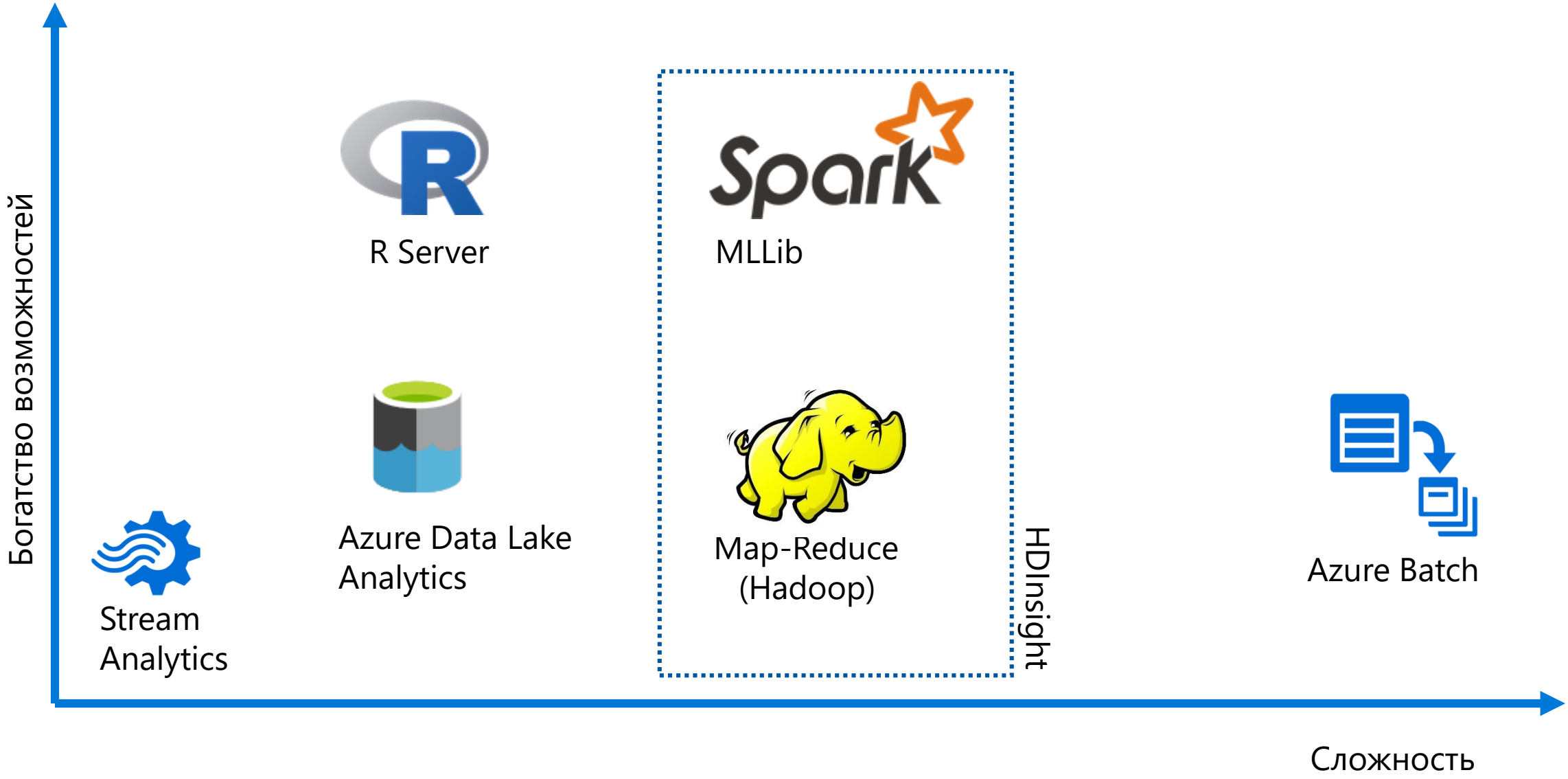
Graph



Global distribution | Elastic scale out | Guaranteed low latency | Five consistency models | Comprehensive SLAs

A globally-distributed, multi-model database service

Обработка данных

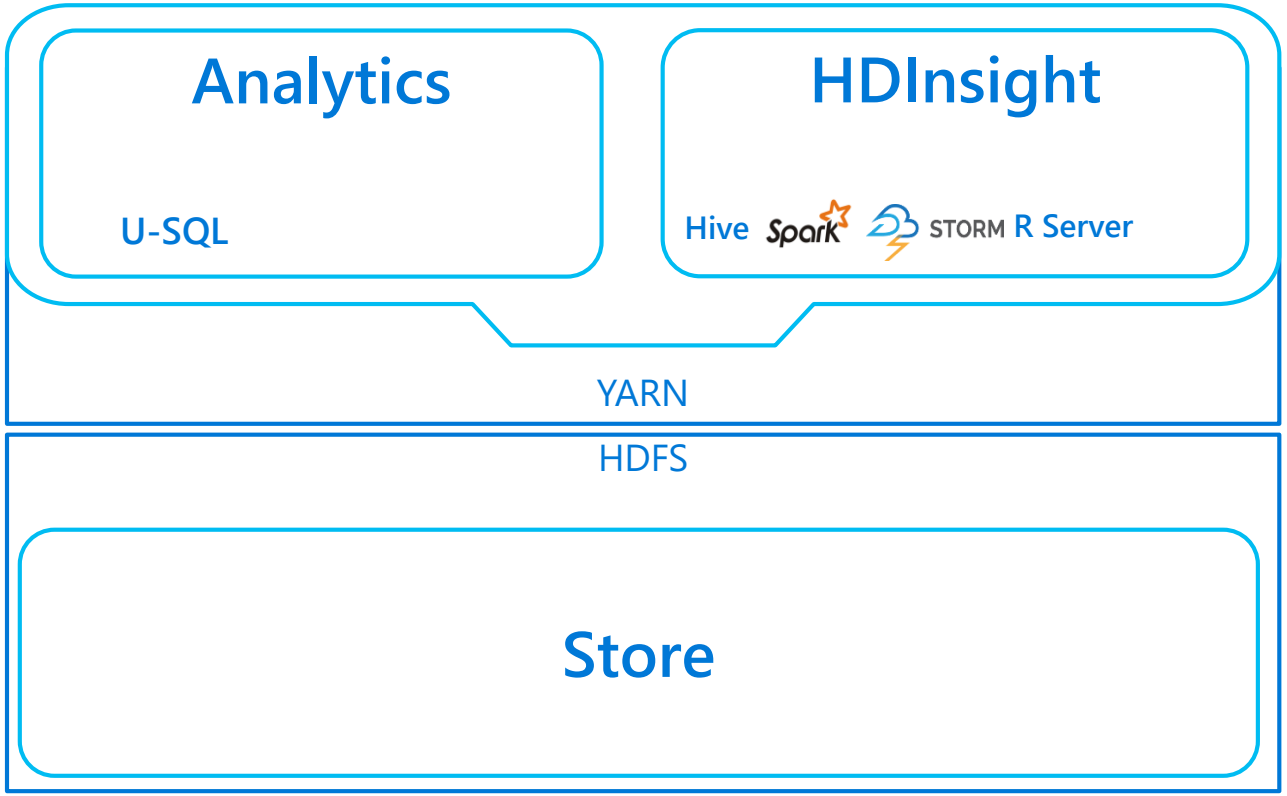


Azure Data Lake

Не думаем об инфрастуктуре

Оплата за выполнение запросов

SQL-подобный язык



Контроль над инфраструкту-рой

Широко используемые открытые технологии

Демонстрация

Обрабатываем тексты

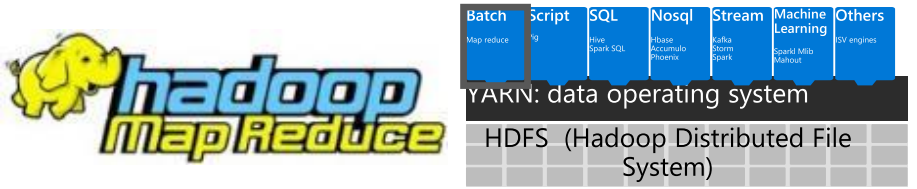
http://gutenberg.org -> Azure Data Lake Analytics

Azure Data Lake Analytics

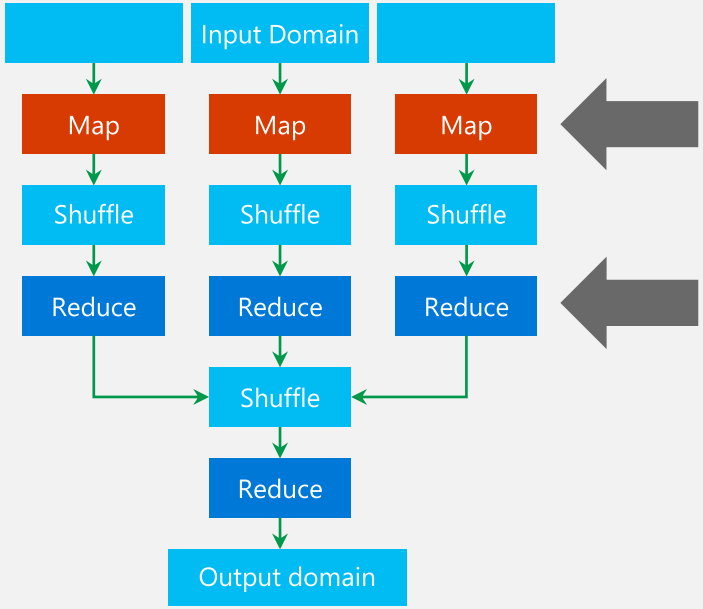
```
@t = EXTRACT date string
           , time string
           , author string
           , tweet string
     FROM "/Input/MyTwitterHistory.csv"
     USING Extractors.Csv();

@res = SELECT author AS author
            , COUNT(*) AS tweetcount
       FROM @t
       GROUP BY author;

OUTPUT @res TO "/Output/MyTwitterAnalysis.csv"
ORDER BY tweetcount DESC
USING Outputters.Csv();
```

- U-SQL: простой и расширяемый язык запросов
- Декларативная природа SQL с мощью C#
- Использование библиотек .NET, R и Python
- Параллелизация кода для различных задач (ETL, ML, image tagging, facial detection)

MapReduce

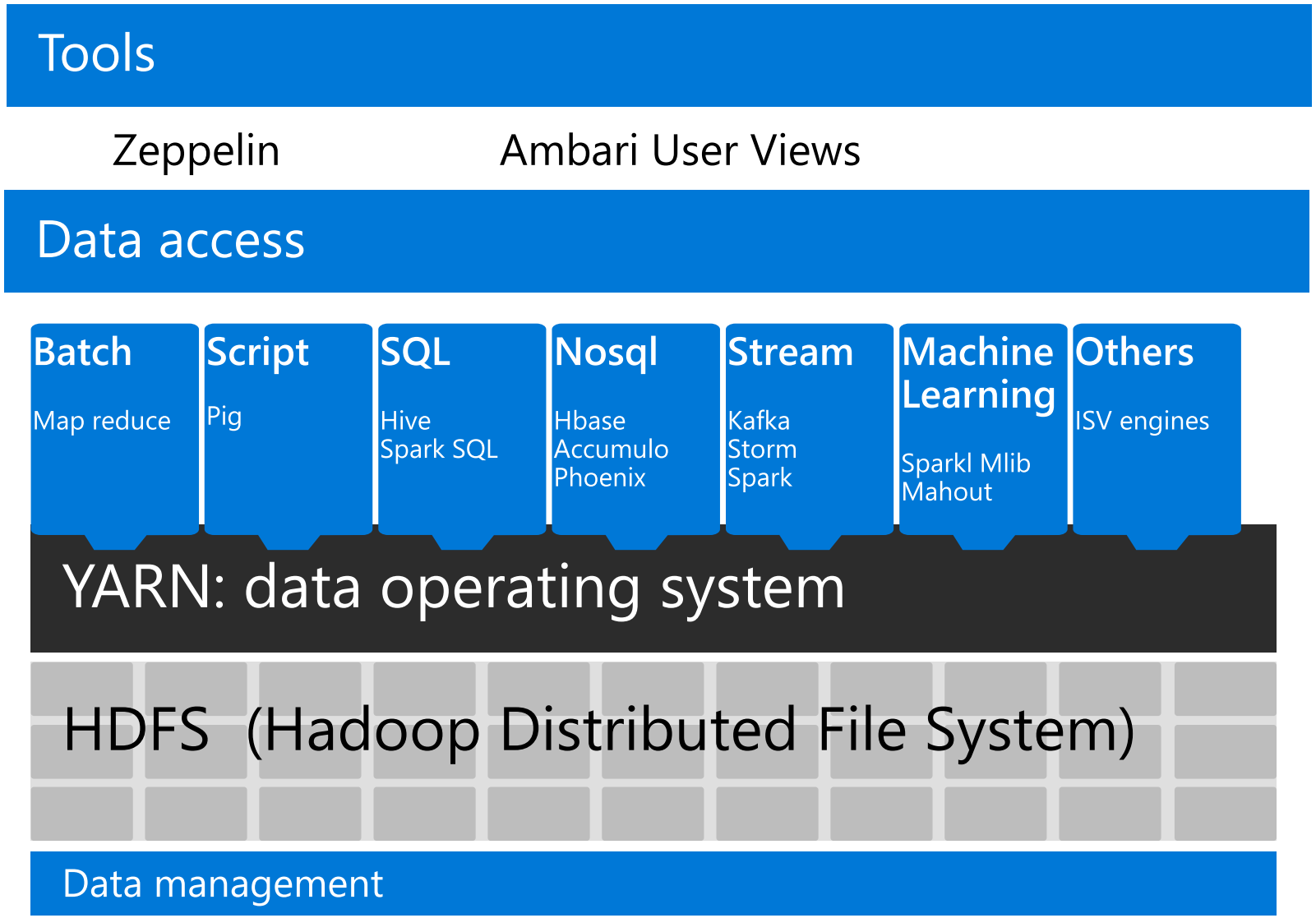


- Обработка происходит там, где хранятся данные
  - Данные хранятся распределенно (HDFS)
  - После первичных вычислений агрегируются только результаты
  - Линейное масштабирование при добавлении узлов
- Фреймворк берет на себя распределение функций по узлам




```
var map = function (key, value, context) {
var words = value.split(/[^a-zA-Z]/);
for (var i = 0; i < words.length; i++) {
          if (words[i] !== "")
context.write(words[i].toLowerCase(),
1);}
}};
var reduce = function (key, values, context) {
var sum = 0;
while (values.hasNext()) {
sum += parseInt(values.next());
     }
context.write(key, sum);
};
```

Hadoop / HDInsight



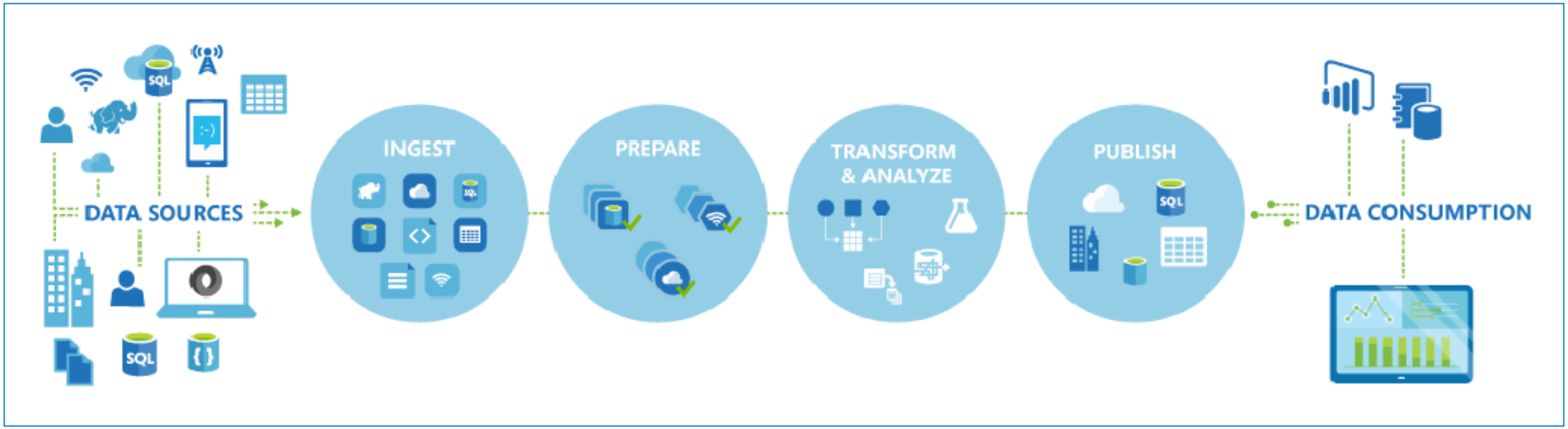
Демонстрация

Обрабатываем тексты

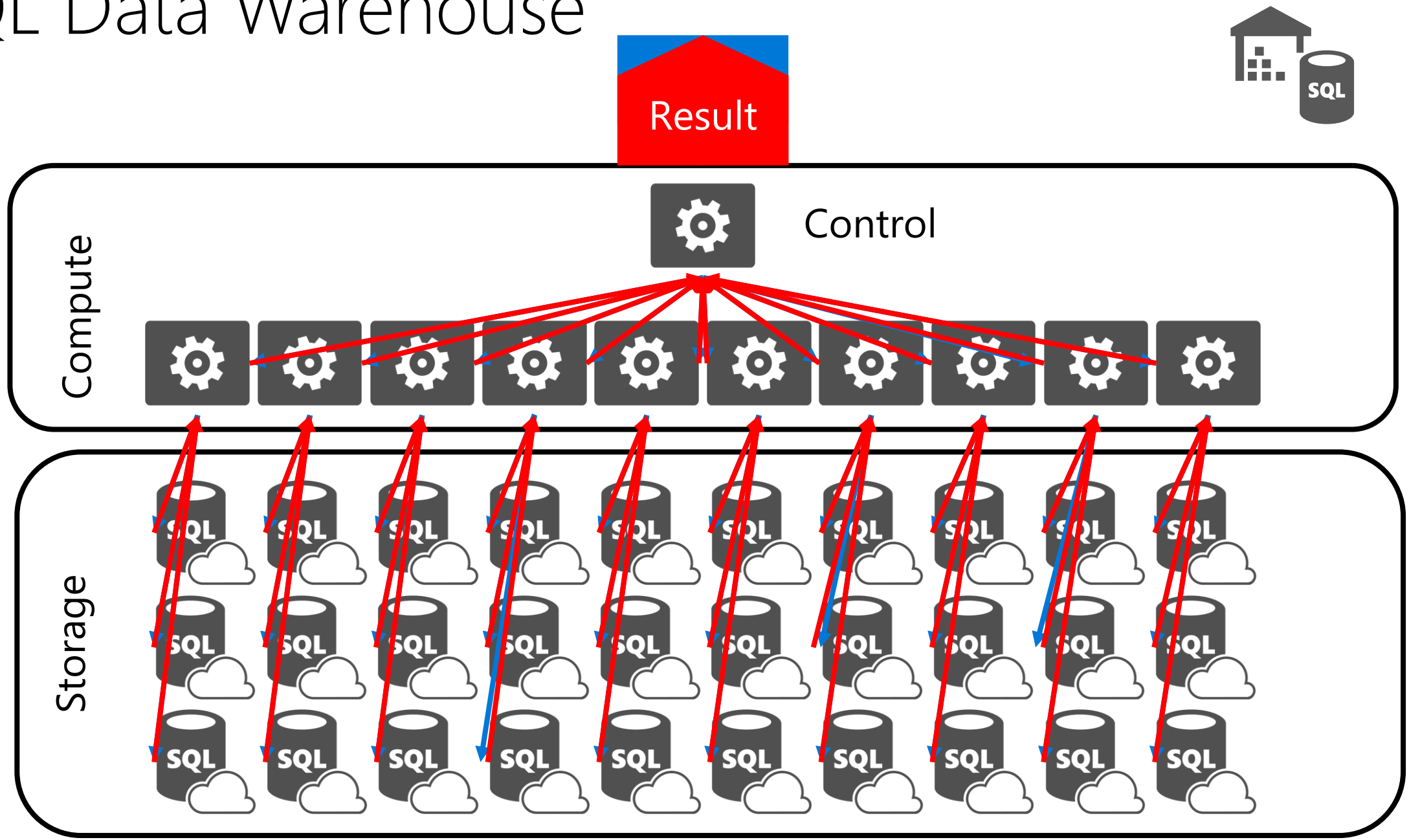
http://gutenberg.org -> Spark

Azure Data Factory

- Оркестрирование всех задач обработки данных
- Pipelines, которые запускаются по расписанию
- Различные источники данных на входе и на выходе
- Внутри pipeline – запросы на U-SQL, Spark/Hadoop и т.д.



SQL Data Warehouse



Простой пример

```
SELECT   COUNT_BIG(*)
FROM     dbo.[FactInternetSales]
;
```

```
SELECT   SUM(*)
FROM     dbo.[FactInternetSales]
;
```

Control

Compute

```
SELECT   COUNT_BIG(*)
FROM     dbo.[FactInternetSales]
;
```

```
SELECT   COUNT_BIG(*)
FROM     dbo.[FactInternetSales]
;
```

```
SELECT   COUNT_BIG(*)
FROM     dbo.[FactInternetSales]
;
```

```
SELECT   COUNT_BIG(*)
FROM     dbo.[FactInternetSales]
;
```

Задачи для SQL Warehouse

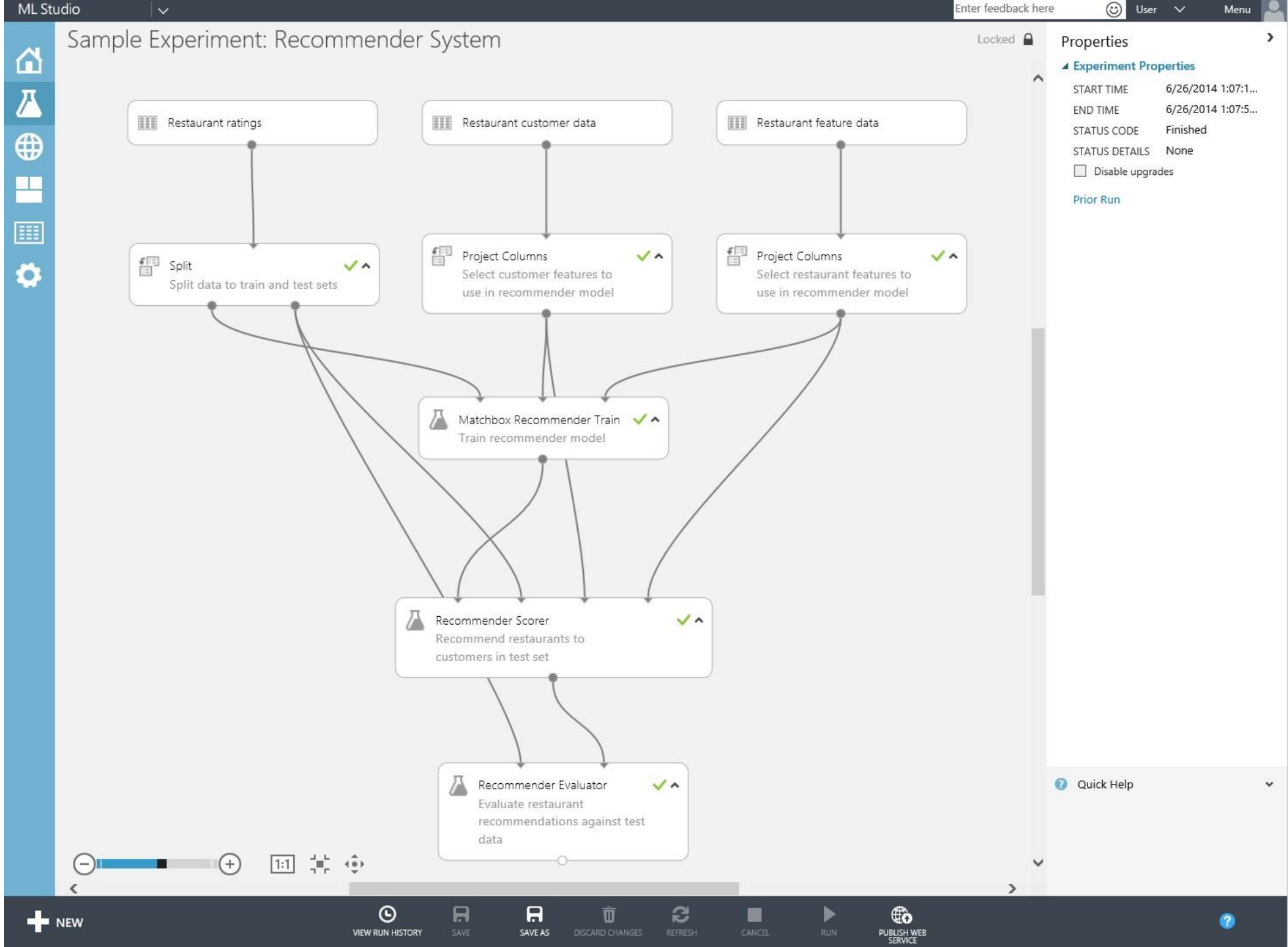
Хорошо подходит

- Хранение больших объемов данных
- Отчеты по большим объемам данных
- Агрегирование данных

Плохо подходит

- Подготовка данных (построчная обработка)
- OLTP
- Большое количество индивидуальных запросов

AZURE ML STUDIO



Эксперты

Евгений Григоренко

Главный эксперт
по Big Data

Ольга Тележная

Python Guru

Дмитрий Сошников

Обработка данных, F#
IoT, ML, CNTK

Jan Pospisil

Приглашенный эксперт

Сегодня: Azure и Big Data (в астрофизике)

Сергей Герасимов

ВМК МГУ

Александр Мещеряков

Институт космических исследований РАН

В четверг

Дмитрий Зобнин

Azure Batch: Big Compute без головной боли (или почти без)

Jan Pospisil

Потоковые технологии обработки больших данных на реальных примерах

В субботу

Андрей Устюжанин

Большие данные в большом адронном коллайдере

Руководитель лаборатории больших данных ФКН НИУ ВШЭ
Руководитель совместных проектов Яндекс-CERN