СБЕРБАНК
Всегда рядом

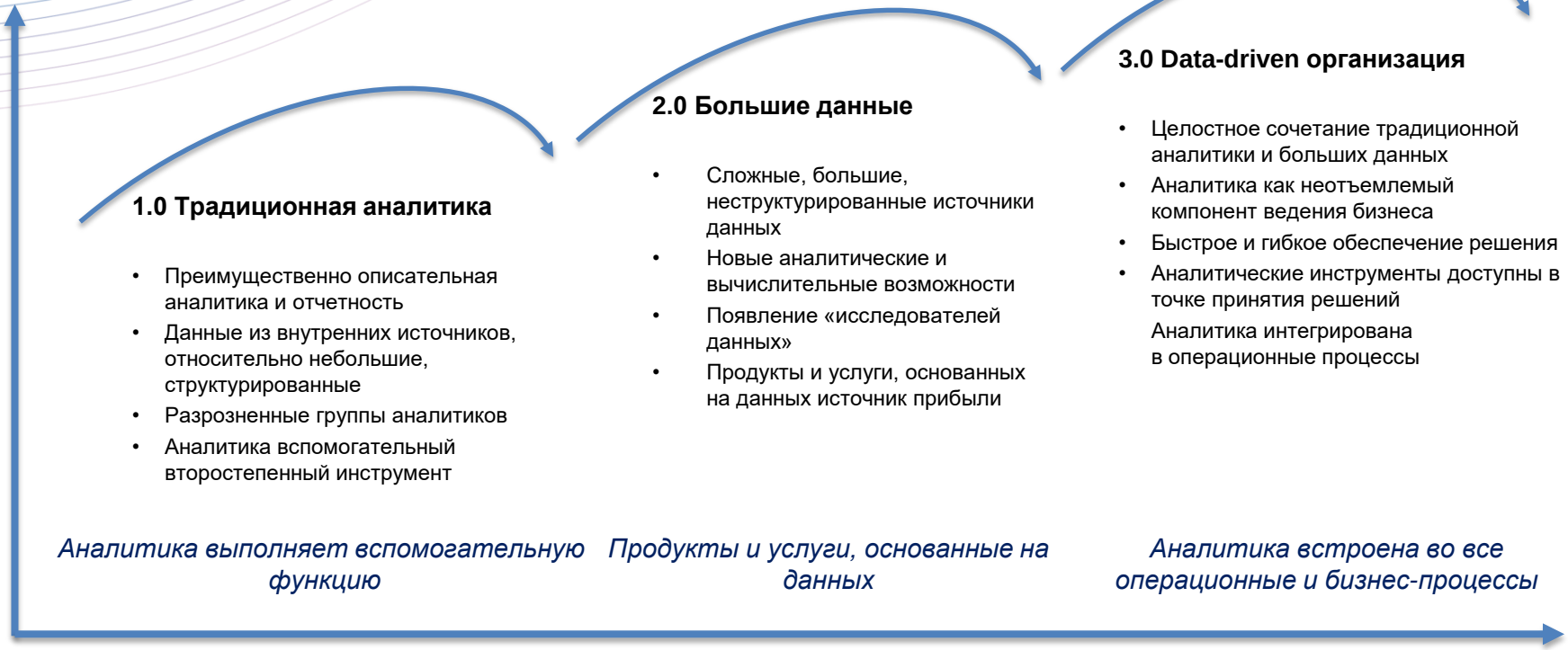
# Развитие DataLab в Сбербанке

Докладчики: Ерофеев А.С, Негго Н.В, Панфилова Д.А.

СБЕРБАНК
Всегда рядом

# Содержание

# Немного теории про эволюцию аналитики и организаций



### 1.0 Традиционная аналитика

- Преимущественно описательная аналитика и отчетность
- Данные из внутренних источников, относительно небольшие, структурированные
- Разрозненные группы аналитиков
- Аналитика вспомогательный второстепенный инструмент

*Аналитика выполняет вспомогательную функцию*

### 2.0 Большие данные

- Сложные, большие, неструктурированные источники данных
- Новые аналитические и вычислительные возможности
- Появление «исследователей данных»
- Продукты и услуги, основанных на данных источник прибыли

*Продукты и услуги, основанные на данных*

### 3.0 Data-driven организация

- Целостное сочетание традиционной аналитики и больших данных
- Аналитика как неотъемлемый компонент ведения бизнеса
- Быстрое и гибкое обеспечение решения
- Аналитические инструменты доступны в точке принятия решений

Аналитика интегрирована в операционные процессы

*Аналитика встроена во все операционные и бизнес-процессы*

# Немного истории развития направления по работе с данными в Сбербанке

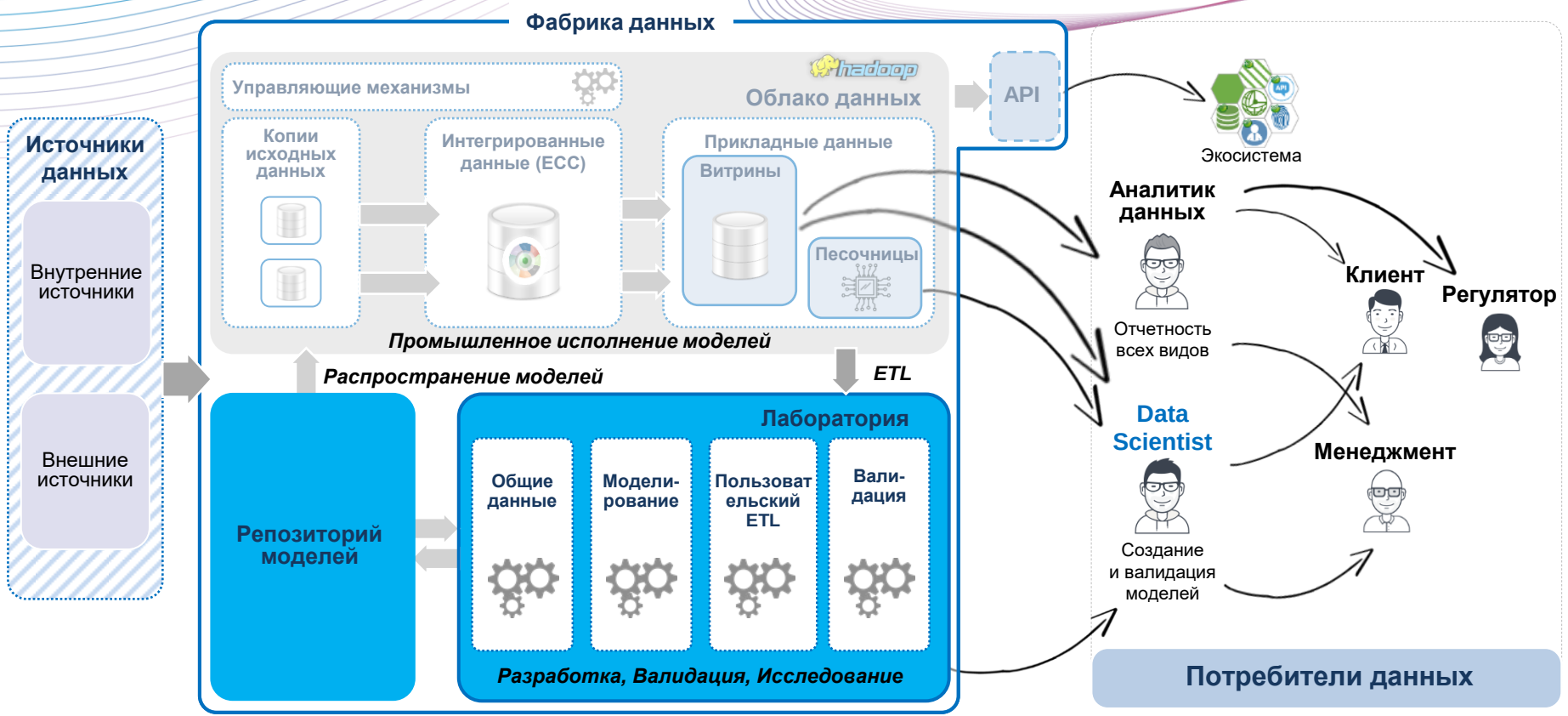

## GAP к внутреннему спросу

- Ассортимент данных и качество данных
- Экспоненциальных рост данных и стоимости их хранения и обработки
- Разрозненные процессы и практики работы с данными в разных бизнес юнитах и командах

## GAP к мировым практикам

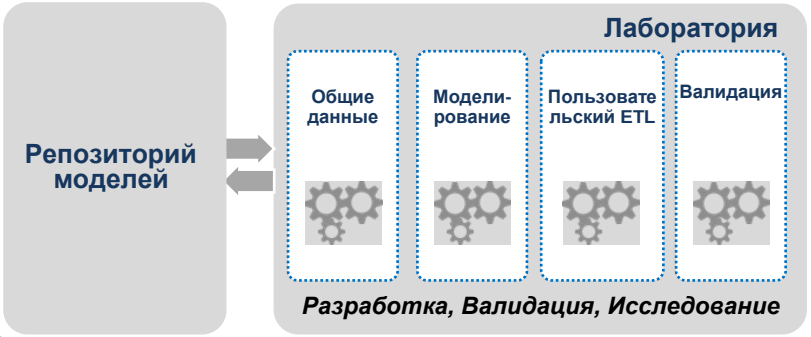
- Начальные уровень зрелости функции Data Governance
- Time to market продуктов на основе данных

# Фабрика данных



# Ключевые системы Фабрики Данных


1) Единое место сбора, хранения и распространения данных
2) Среда промышленного исполнения аналитических моделей




1) Исследование данных
2) Поиск знаний и проверка гипотез
3) Разработка и валидация моделей

# Data Lab

# К нам пришли Data Scientists!



Современное ПО

Удобные ИТ-сервисы

Data Lake

Мощное железо

Спиннер

Смузи и печеньки

# Мы выбрали единую Лабораторию данных для Банка



СБЕРБАНК
Всегда рядом

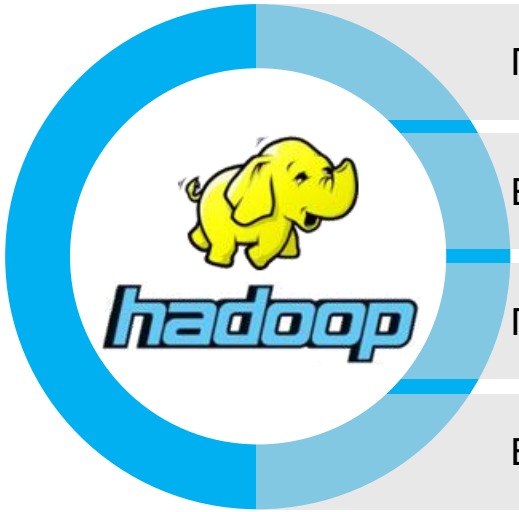
## Вызовы

- Отсутствие возможности горизонтального масштабирования
- Архитектура повторяет орг. структура - дублирование систем и расходов на сопровождение и развитие
- Разделены системы хранения данных и аналитики
- Ограничения по объемам данных для аналитики и машинного обучение

## Наше решение – единая Лаборатория для всех функциональных блоков

- Снижение затрат на администрирование и сопровождение
- Единые инсталляции продуктов на всю платформу
- Данные собраны и доступны в единой системе и публикуются для всех
- Общие стандарты и инструменты для безопасности данных

# Лабораторию построили на основе стека Hadoop




- Подходит для машинного обучения
- Есть возможность разработки на Python, R, Scala, Hive, Impala и др.
- Позволяет хранить и анализировать большие объемы данных
- Возможность использования Open Source решений

# Платформа состоит из…

**SQL аналитик на данных Hadoop**


**Вычисления в памяти**


**Языки**


**ETL**

Informatica Big Data Management

**Доставка данных**

Sqoop


**Управление ресурсам кластера**

**YARN**

**Распределенная файловая система кластера**


**Приложения**



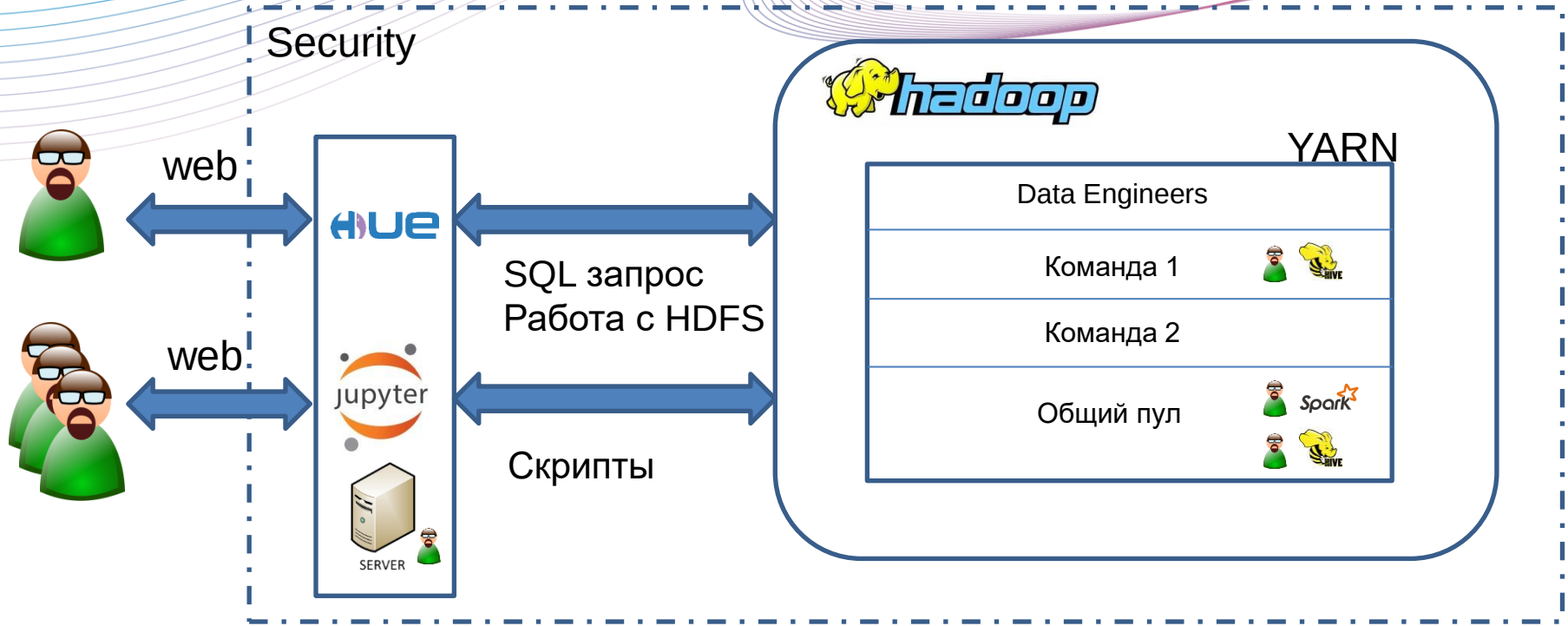
**Security**

Kerberos
Sentry

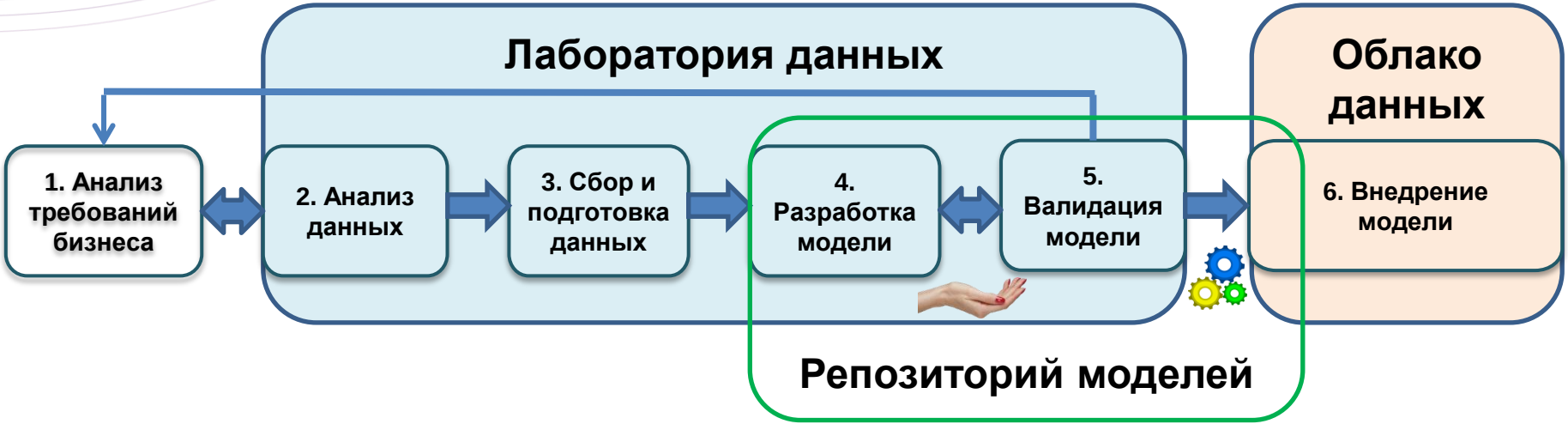
**OpenSource** – ПО, распространяемое с открытым исходным кодом, как свободное для изучения, переиспользования и модификации исходного кода конечными пользователями

# Физика работы на примере четырех пользователей-DS





# Жизненный цикл исследования данных




**Лаборатория данных**

- Разработка модели
- Валидация модели

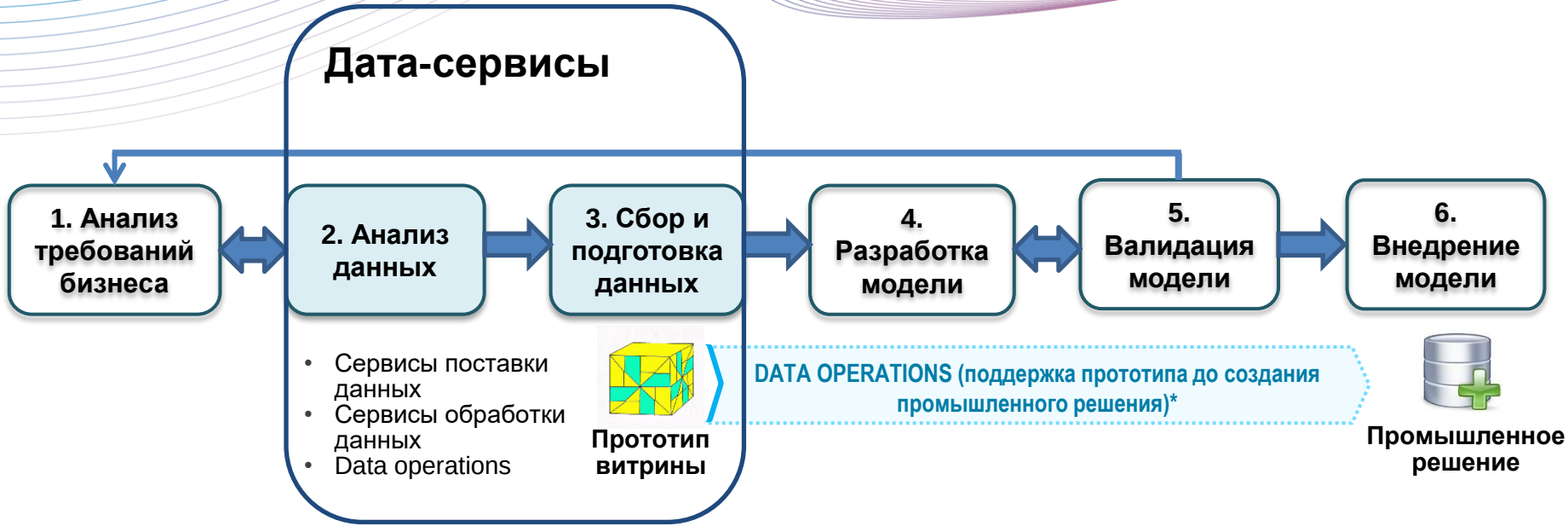
**Репозиторий моделей**

- Хранение кода, описания и метаданных модели

**Облако данных**

- Промышленное исполнение модели

# Жизненный цикл исследования данных




* Есть ограничение по срокам предоставления сервиса

# Дата-сервисы

## Нальем данных!

### Поставка данных

- Поставка данных в Лабораторию данных в рамках SLA
- Единое информационное пространство управления метаданными (реестр и каталог данных в Карте данных)

## Соберем датасет!

### Обработка данных

- Сервисы
  - Дополнительная обработка данных по требованию заказчика
  - Проверка качества данных
  - Интеграция данных
  - Построение агрегатов
- Формируется бэклог задач, который находится в едином информационном пространстве (Jira)
- Задачи исполняются по FIFO

## Поддержим прототип!

### Data Operations

- Заключение соглашения о поддержке прототипа решения на фиксированный период времени
- Возможность использования прототипа решения до его внедрения на ПРОМ
- Выделенные ресурсы на поддержку прототипа

Сервисы предоставляются выделенной группой Data Engineers

# Что у нас есть сейчас



- Работают более 300 Data Scientists
- Доступны более 50 систем-источников
- Более 200 Open Source библиотек для работы с Python и  более 1000 для работы с R
- Решение SAS HPA

# Дальнейшие шаги





1. Создание и запуск моделей на real-time потоках данных
2. Использование GPU для обучения моделей с использованием нейронных сетей
3. Дообучение моделей на промышленных потоках данных