СБЕРБАНК
Всегда рядом

Подходы к организации работ с большими данными в Сбербанке

Москва, 19 апреля 2018

Новая платформа Банка

СБЕРБАНК
Всегда рядом

Единая Фронтальная Система

ЕФС

Единое «Окно» контакта с клиентом

2018+

БИЗНЕС-ХАБ

Профиль Клиента

«Интеллект» - логика и система принятия решений

ПРОДУКТОВЫЕ ФАБРИКИ

«Производство» - продукты, транзакции, учет

ФАБРИКА ДАННЫХ

Исторический профиль Клиента

«Опыт» - накопление клиентского опыта и генерация новых знаний

Трансформация в data driven модель

СБЕРБАНК
Всегда рядом

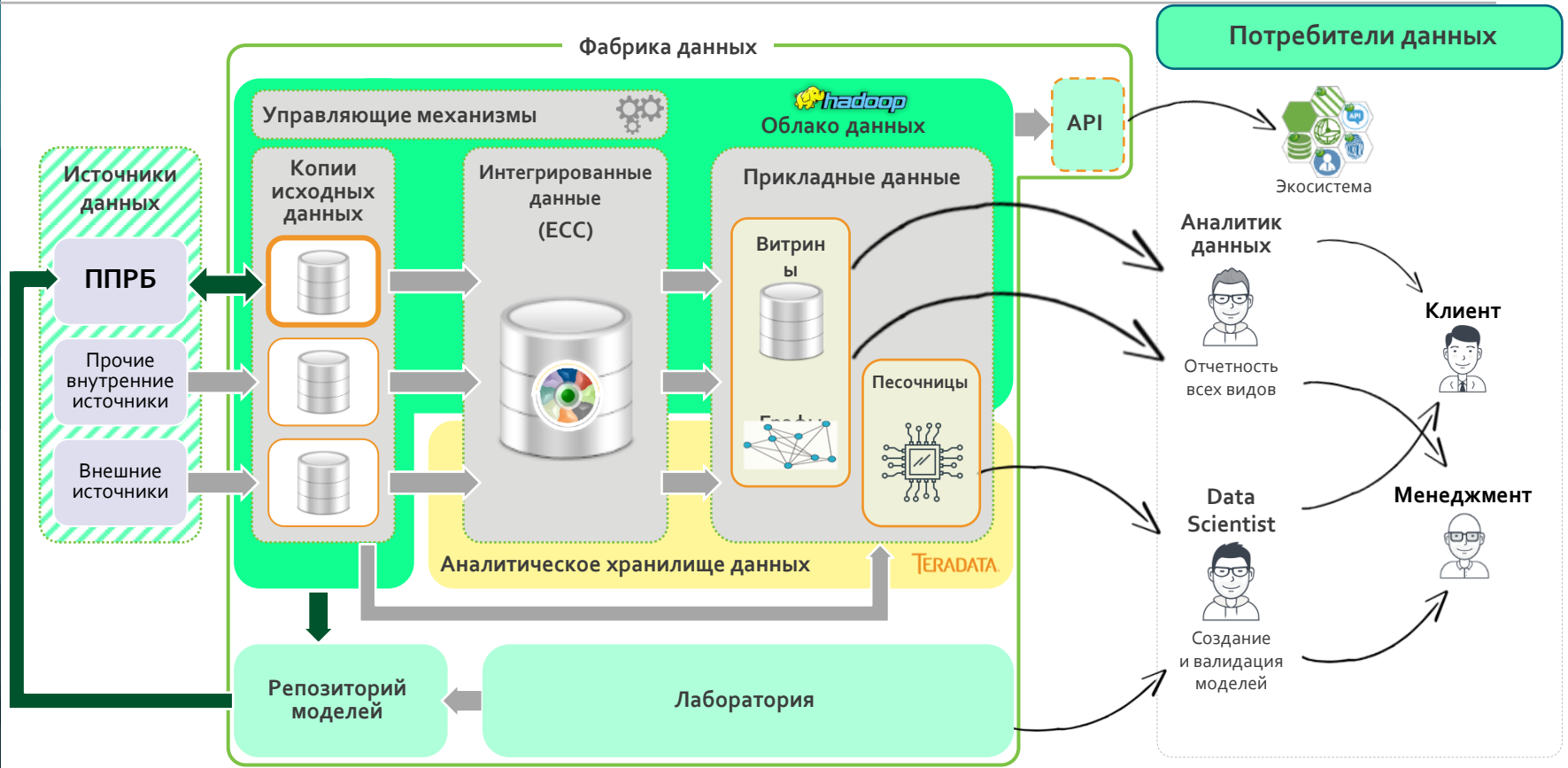
- За 2017 год создана новая инфраструктура, которая позволяет сегодня решать принципиально новые задачи, чтобы уже в ближайшем будущем встать в один ряд с технологическими компаниями

- 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

- Данные – актив, которым нужно управлять: Каталог, Качество, Доступность
- Новые роли и компетенции – Data Scientist, Data Engineer
- Новые подходы к решению задач

Архитектура Фабрики данных

СБЕРБАНК
Всегда рядом

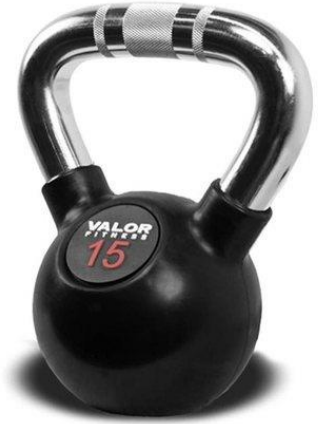


Облако данных. Факты

СБЕРБАНК
Всегда рядом

Железо:

- 250 узлов в составе кластера
- 170 узлов Hadoop data node (HDFS + YARN):
  - ✓ 6800 CPU Cores
  - ✓ 109 Тб оперативной памяти
  - ✓ 10 Пб HDD



Данные:

- 2.5 Пб - объем данных, план – рост до 10 ПБ в течение 2018г.
- 15.5 Тб в сутки – объем получаемых изменений
- 2000 – 5000 транзакций в секунду
- 400 параллельных заданий загрузки данных
- 200 ТБ - объем ежесуточно обновляемой информации в репликах
- 6 часов - время на ежесуточное обновление данных

Использованные технологии

СБЕРБАНК
Всегда рядом

SQL аналитика на данных Hadoop

Вычисления в памяти

NoSQL СУБД

ETL

Индексный поиск

Интеграция и потоковая обработка

Администр.

Управление

Координация

Управление ресурсам кластера

YARN

Распределенная файловая система кластера

ECC и Презентационные слои

СБЕРБАНК
Всегда рядом

Единый семантический слой (ECC):

- Единая версия правды
- В основе - Клиентоцентрическая модель сущностей
- Является основным источником для потребителей данных
- Хранит данные в Parquet формате
- Может обновляться batch и streaming процессами

Специализированные витрины данных:

- Рассчитываются на основе ECC и прочих источников
- Расчет только batch процессами
- Всегда конкретный заказчик
- Могут являться входом и выходом какого-либо процесса
- Данные хранятся в формате Parquet

Работа с моделями

СБЕРБАНК
Всегда рядом

ФП Библиотека Моделей:

- ❑ Хранение кода в git репозитории
- ❑ Версионирование
- ❑ Передача кода на исполнение

ФП Пакетное Исполнение Моделей:

- ❑ Поддержка моделей на SAS, Python, Scala, Java
- ❑ Извлечение моделей из Библиотеки
- ❑ Исполнение моделей под Spark
- ❑ Сохранение результатов работы модели

ML, AI. Применение в банке

СБЕРБАНК
Всегда рядом

	Дескриптивная аналитика	Предиктивная аналитика	Предписывающая аналитика
	Что происходит?	**Что дальше?**	**Как мы можем повлиять?**
Описание	Выделение ключевых характеристик, группировка данных	Прогнозирование вероятности наступления будущих событий	Рекомендация управляющих действий
Принцип			
Примеры	Сегментация клиентов Классификация типов событий	Прогноз показателей Прогноз банкротства Предсказание надежности клиента	Персонализация предложений Блокировка мошеннических транзакций

Спасибо за внимание

СБЕРБАНК
Всегда рядом

Борис Рабинович

Сбербанк-Технологии

Директор центра компетенции развития BI технологий

BIRabinovich.sbt@sberbank.ru