

Создание витрин данных и интеграционных процессов на платформе Hadoop

Апрель 2023

Наша цель с точки зрения бизнеса

Дать каждому клиенту почувствовать себя уникальным при взаимодействии со «Сбером».



Это значит, мы должны понимать какая следующая потребность возникнет у клиента, чтобы предложить лучшее решение этой потребности в виде кастомизированных продуктов/сервисов Сбера и компаний партнёров.

Техническая задача

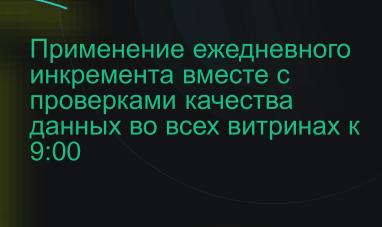
Сформировать витрины данных и запустить интеграционные процессы, которые позволят выполнять задачи по запуску коммуникаций, построению математических моделей, аналитики, отчётности

сырых реплик источников данныхбщий объем реплик

100+

15⁺Птб 10⁺тб

ежедневный инкремент



Архитектура решения

Система

Система

источник N

источник 1



Витрины данных

GREENPLUM

Phadoop

Более активных пользователей ежемесячно

Система приемник 1

Система приемник N

Система отчетности и

аналитики

Сырые

данные *Phadoop*

100+ реплик

источников общим

объемом

234 реализованные

витрины, обновляемые в режиме Batch

16 NRТ-потоков обработки данных

Текущий объем, занимаемый витринами -

9,5 ⊓тб

Spark











10+Tb данных ежедневный инкремент

















Инструменты контроля качества данных + АІ-инструменты КД

Что получили в итоге?

Обработка 10+ТВ инкремента данных с применением сложных аналитических функций к ним за 4-5 часов



Перевод части кампаний, которые ранее использовали данные с актуальностью Т-1, на NRT платформу, что позволило на 12% сократить кол-во жалоб и увеличить NPV на 20%

Спасибо

Илья Нечаев

Исполнительный директор Департамента данных и рекомендательных систем в СберБанке