Введение в BigData

11.02.2025

Содержание

01

Немного истории

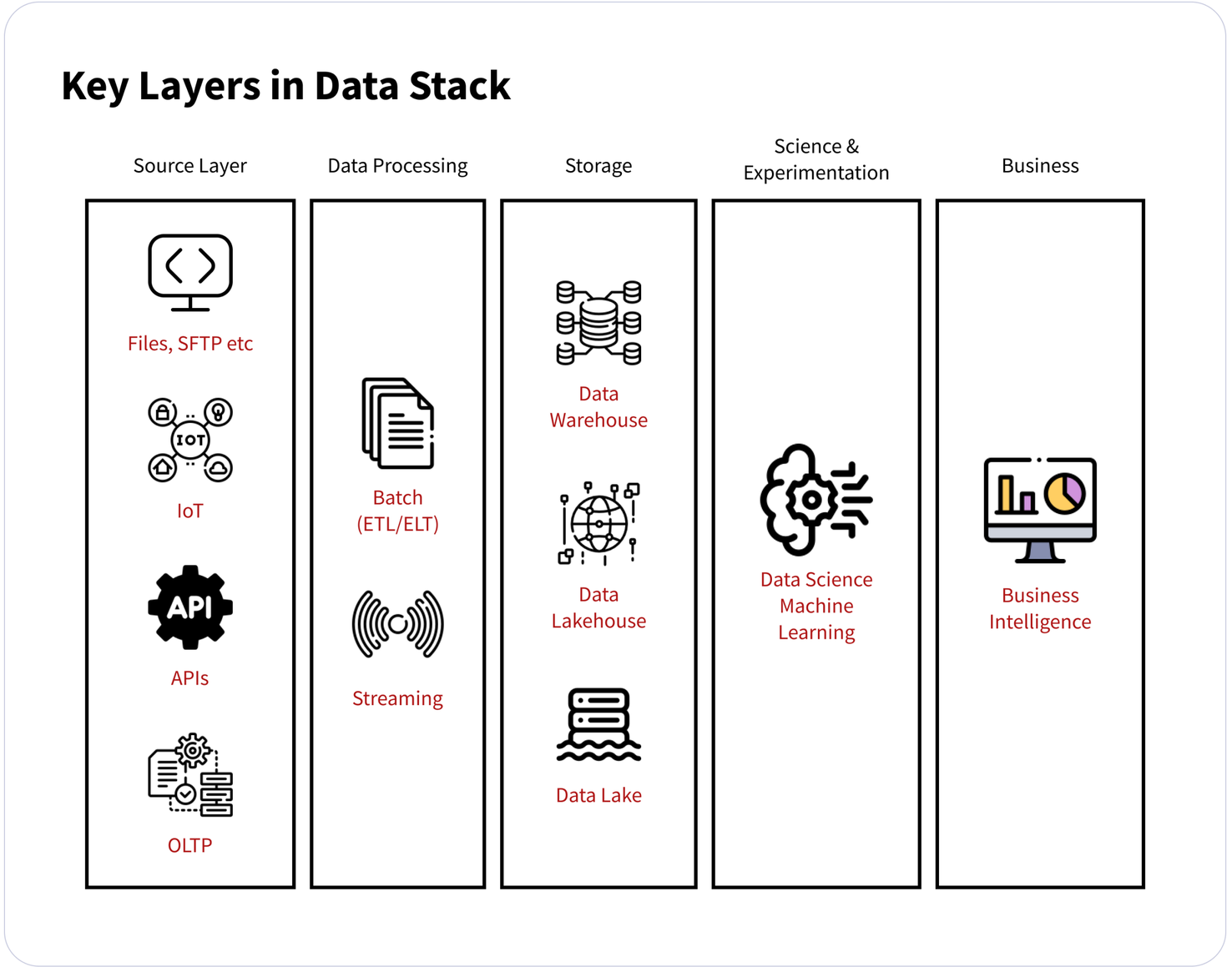
Немного истории

1. **70-ые–** Эдгар Кодд появления реляционной модели данных;
2. **80-ые –** Пол Мерфи, Барри Девлин “Business DWH”;
3. **90-ые –** Появления Oracle, Teradata, IBM D2. Модели Билла Инмона и Ральфа Кимбалла;
4. **00-ые –** MPP субд, появление Greenplum, Hadoop; Термин BigData(Клиффорд Линч)
5. **10-ые** – Эпоха облаков (Snowflake, G BigQuery, Amazon Redshift)
6. **20-ые** – Trino, YugaByte и новые.

02

Архитектура решений

Архитектура решений



Source Layer

1. **Files, SFTP, etc —** файловые ресурсы(Excel, csv, FTP|SFTP);
2. **IOT -** internet of things, устройства умного дома, медицинское оборудование и т.д;
3. **API -** Application programming interface, программный интерфейс для получения данных;
4. **OLTP —** Online Transaction Processing. Системы обработки транзакций в реальном времени.

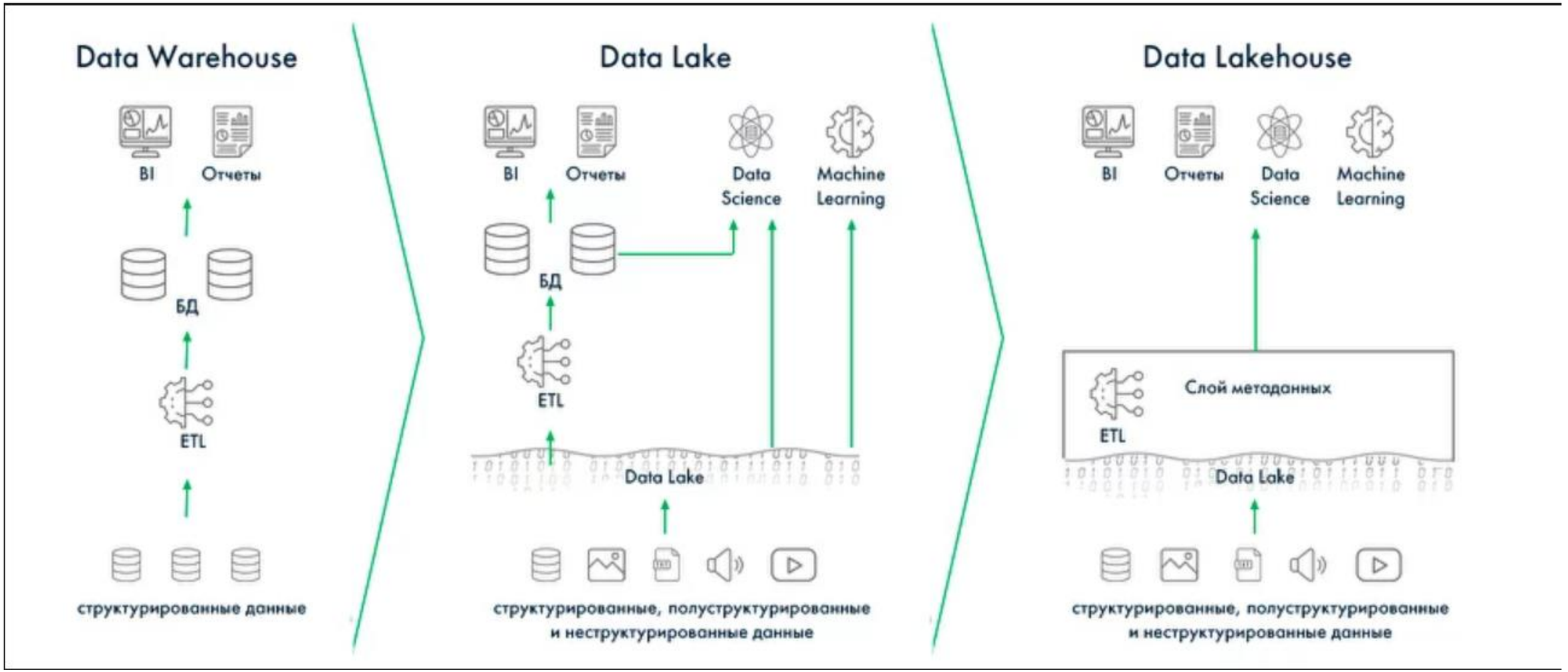
Data processing

1. **Batch —** ELT, ETL(Extract, Load, Transform и Extract, Transform, Load). Системы передачи **порционных** данных сырых данных с источника.
2. **Streaming** — обработка данных в реальном времени.

Storage

1. **DWH —** Хранилище данных. Это система для хранения и управления большими объемами структурированных данных, которые объединяются из различных источников с целью получения из них бизнес выгоды (G BigQuerry, Amazon Redshift, Snowflake, Greenplum).
2. **Data Lake —** Озеро данных. Это инструмент хранения необработанных(неструктурированных) данных в различных форматах (Amazon S3, Minio, G Cloud Storage, Yandex S3).
3. **Data Lakehouse —** Объединение концепции DWH и Data Lake. Храним все виды данных и управляем метаданными.

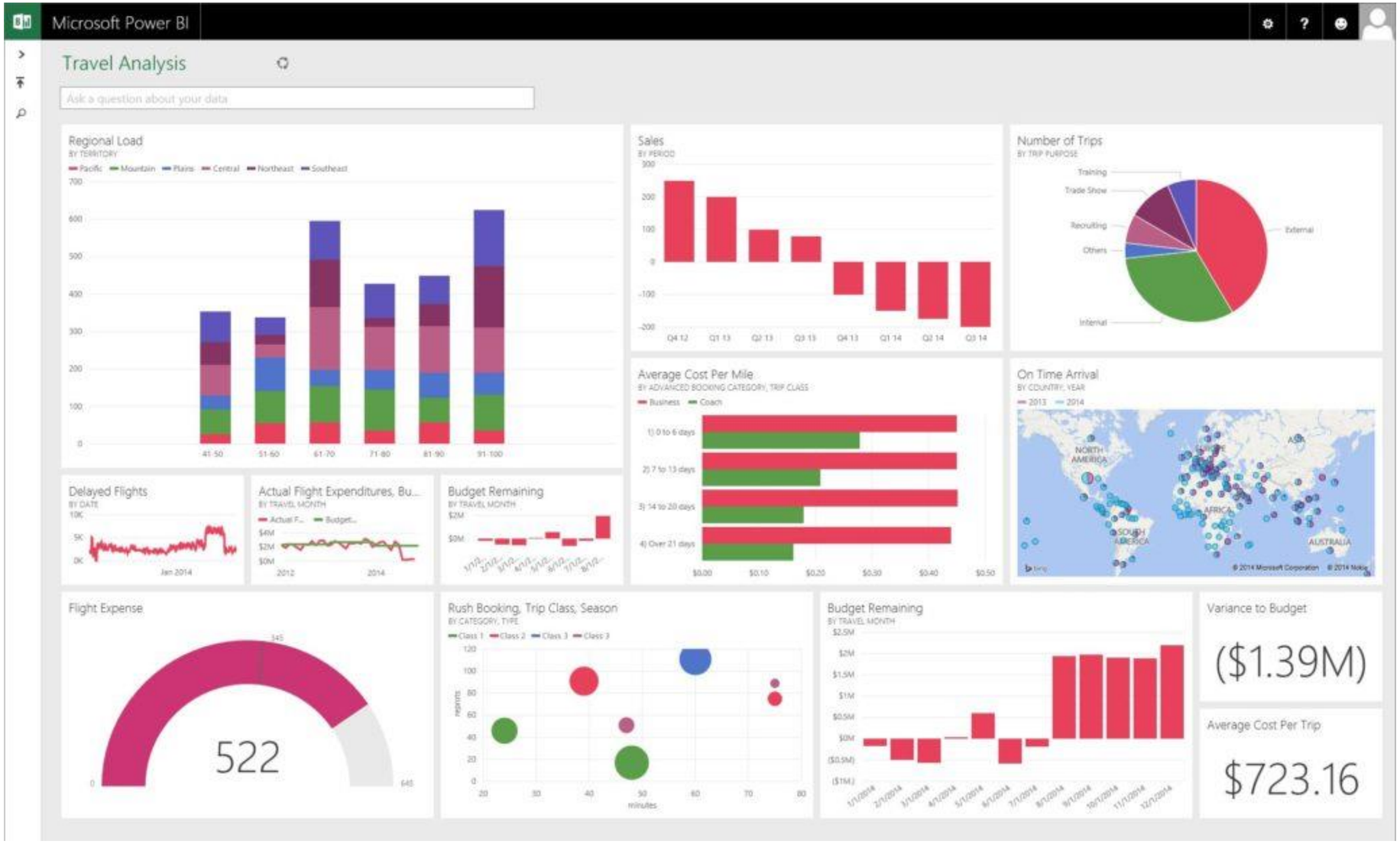
Storage


Data Warehouse
BI
Отчеты
БД
ETL
структурированные данные
Data Lake
BI
Отчеты
Data Science
Machine Learning
БД
ETL
Data Lake
структурированные, полуструктурированные и неструктурированные данные
Data Lakehouse
BI
Отчеты
Data Science
Machine Learning
Слой метаданных
ETL
Data Lake
структурированные, полуструктурированные и неструктурированные данные

Science&Experimentation

1. **Adhoc —** запросы к источнику данных для получения метрик, подтверждения гипотез;
2. Скрипты, аналитика и т.д.
3. **ML** модели.

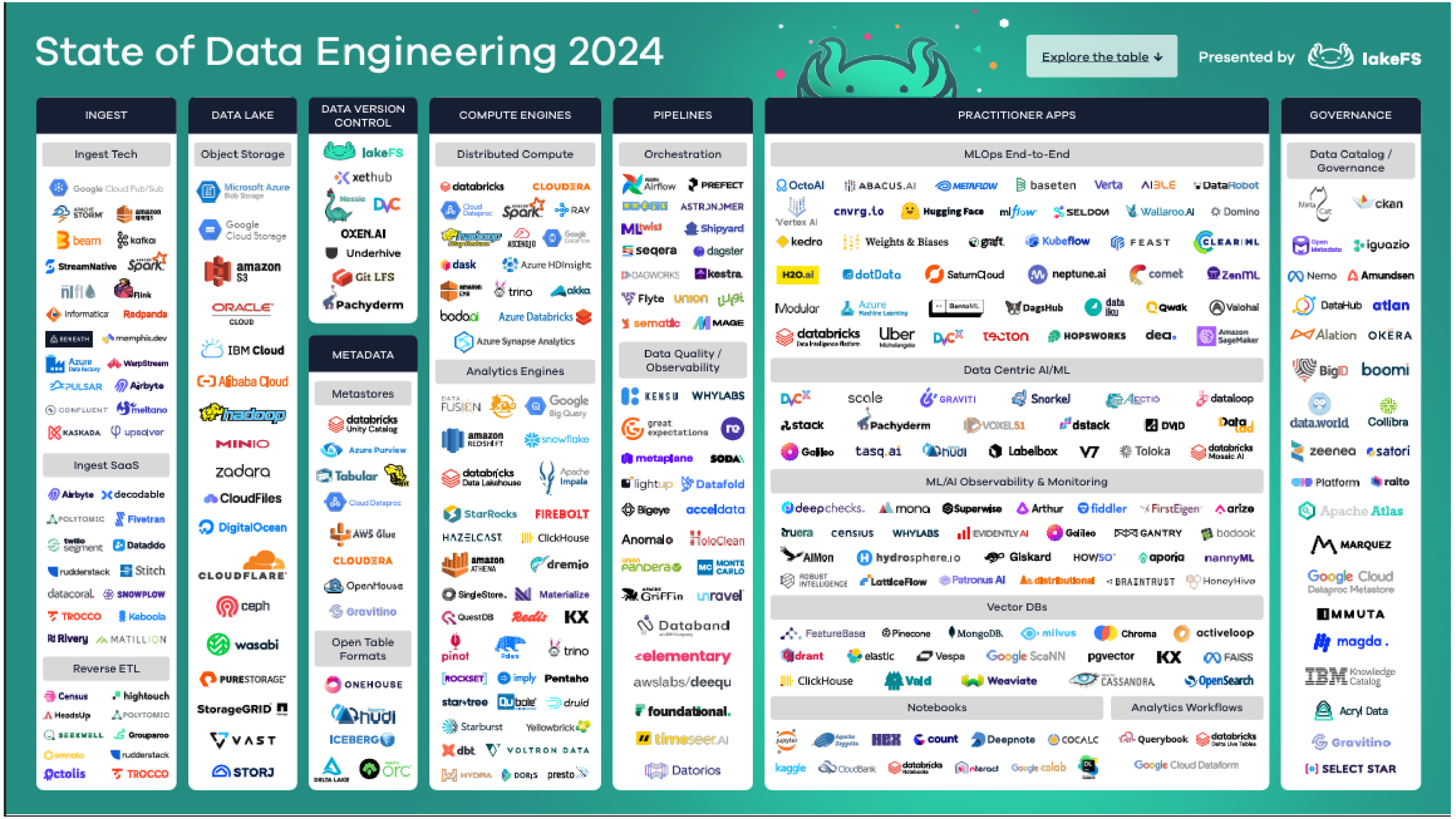
BI


Microsoft Power BI
Travel Analysis
Ask a question about your data
Regional Load
BY TERRITORY
Pacific
Mountain
Plains
Central
Northeast
Southeast
41-50
51-60
61-70
71-80
81-90
91-100
Sales
BY PERIOD
Q4 12
Q1 13
Q2 13
Q3 13
Q4 13
Q1 14
Q2 14
Q3 14
Number of Trips
BY TRIP PURPOSE
Training
Trade Show
Recruiting
Others
Internal
External
Average Cost Per Mile
BY ADVANCED BOOKING CATEGORY, TRIP CLASS
Business
Coach
1) 0 to 6 days
2) 7 to 13 days
3) 14 to 20 days
4) Over 21 days
On Time Arrival
BY COUNTRY, YEAR
2013
2014
NORTH AMERICA
ASIA
AFRICA
SOUTH AMERICA
AUSTRALIA
Delayed Flights
BY DATE
Jan 2014
Actual Flight Expenditures, Bu...
BY TRAVEL MONTH
Budget Remaining
BY TRAVEL MONTH
Flight Expense
522
Rush Booking, Trip Class, Season
BY CATEGORY, TYPE
Class 1
Class 2
Class 3
minutes
Budget Remaining
BY TRAVEL MONTH
Variance to Budget
($1.39M)
Average Cost Per Trip
$723.16

03

Наши дни

Наши дни


State of Data Engineering 2024
Explore the table
Presented by lakeFS
INGEST
Ingest Tech
Google Cloud Pub/Sub
Apache Storm
amazon kinesis
beam
kafka
StreamNative
Spark
nifi
Flink
Informatica
Redpanda
BENEATH
memphis.dev
Azure Data Factory
WarpStream
PULSAR
Airbyte
CONFLUENT
meltano
KASKADA
upsolver
Ingest SaaS
Airbyte
decodable
POLYTOMIC
Fivetran
twilio segment
Dataddo
rudderstack
Stitch
datacoral
SNOWPLOW
TROCCO
Kaboola
Rivery
MATILLION
Reverse ETL
Census
hightouch
HeadsUp
POLYTOMIC
SEEKWELL
Grouparoo
omnata
rudderstack
Octolis
TROCCO
DATA LAKE
Object Storage
Microsoft Azure Blob Storage
Google Cloud Storage
amazon S3
ORACLE CLOUD
IBM Cloud
Alibaba Cloud
hadoop
MINIO
zadara
CloudFiles
DigitalOcean
CLOUDFLARE
ceph
wasabi
PURESTORAGE
StorageGRID
VAST
STORJ
DATA VERSION CONTROL
lakeFS
xethub
Nessie
DVC
OXEN.AI
Underhive
Git LFS
Pachyderm
METADATA
Metastores
databricks Unity Catalog
Azure Purview
Tabular
Hive
Cloud Dataproc
AWS Glue
CLOUDERA
OpenHouse
Gravitino
Open Table Formats
ONEHOUSE
Apache hudi
ICEBERG
DELTA LAKE
orc
COMPUTE ENGINES
Distributed Compute
databricks
CLOUDERA
Cloud Dataproc
Spark
RAY
hadoop
ASCENDIO
Google Dataflow
dask
Azure HDInsight
amazon EMR
trino
akka
bodo.ai
Azure Databricks
Azure Synapse Analytics
Analytics Engines
DATA FUSION
Google Big Query
amazon REDSHIFT
snowflake
databricks Data Lakehouse
Apache Impala
StarRocks
FIREBOLT
HAZELCAST
ClickHouse
amazon ATHENA
dremio
SingleStore
Materialize
QuestDB
Redis
KX
pinot
Pulsar
trino
ROCKSET
imply
Pentaho
startree
Kyuubi
druid
Starburst
Yellowbrick
dbt
VOLTRON DATA
HYDRA
DORIS
presto
PIPELINES
Orchestration
Airflow
PREFECT
DAGSTER
ASTRONOMER
MLtwist
Shipyard
seqera
dagster
DAGWORKS
kestra
Flyte
UNION
Luigi
sematic
MAGE
Data Quality / Observability
KENSU
WHYLABS
great expectations
re
metaplane
SODA
lightup
Datafold
Bigeye
acceldata
Anomalo
HoloClean
union pandera
MONTE CARLO
GriffinAI
unravel
Databand
elementary
awslabs/deequ
foundational.
timeseer.AI
Datorios
PRACTITIONER APPS
MLOps End-to-End
OctoAI
ABACUS.AI
METAFLOW
baseten
Verta
AIBLE
DataRobot
Vertex AI
cnvrg.io
Hugging Face
mlflow
SELDON
Wallaroo.AI
Domino
kedro
Weights & Biases
graft
Kubeflow
FEAST
CLEARML
H2O.ai
dotData
SaturnCloud
neptune.ai
comet
ZenML
Modular
Azure Machine Learning
DemoML
DagsHub
data iku
Qwak
Valohai
databricks Data Intelligence Platform
Uber Michelangelo
DVC
tecton
HOPSWORKS
deca.
Amazon SageMaker
Data Centric AI/ML
DVC
scale
GRAVITI
Snorkel
AEGIS
dataloop
stack
Pachyderm
VOXEL51
dstack
DVID
Data Toddler
Galileo
tasq.ai
hudi
Labelbox
V7
Toloka
databricks Mosaic AI
ML/AI Observability & Monitoring
deepchecks
mona
Superwise
Arthur
fiddler
FirstEigen
arize
truera
census
WHYLABS
EVIDENTLY AI
Galileo
GANTRY
badook
AIMon
hydrosphere.io
Giskard
HOW2SO
aporia
nannyML
ROBUST INTELLIGENCE
LatticeFlow
Patronus AI
distributional
BRAINTRUST
HoneyHive
Vector DBs
FeatureBase
Pinecone
MongoDB
milvus
Chroma
activeloop
qdrant
elastic
Vespa
Google ScaNN
pgvector
KX
FAISS
ClickHouse
Vald
Weaviate
CASSANDRA
OpenSearch
Notebooks
jupyter
Apache Zeppelin
HEX
count
Deepnote
COCALC
kaggle
CloudBank
databricks Notebooks
nteract
Google colab
Analytics Workflows
Querybook
databricks Delta Live Tables
Google Cloud Dataform
GOVERNANCE
Data Catalog / Governance
Meta Cat
ckan
Open Metadata
iguazio
Nemo
Amundsen
DataHub
atlan
Alation
OKERA
BigID
boomi
data.world
Collibra
zeenea
satori
Platform
raito
Apache Atlas
MARQUEZ
Google Cloud Dataproc Metastore
MMUTA
magda.
IBM Knowledge Catalog
Acryl Data
Gravitino
SELECT STAR

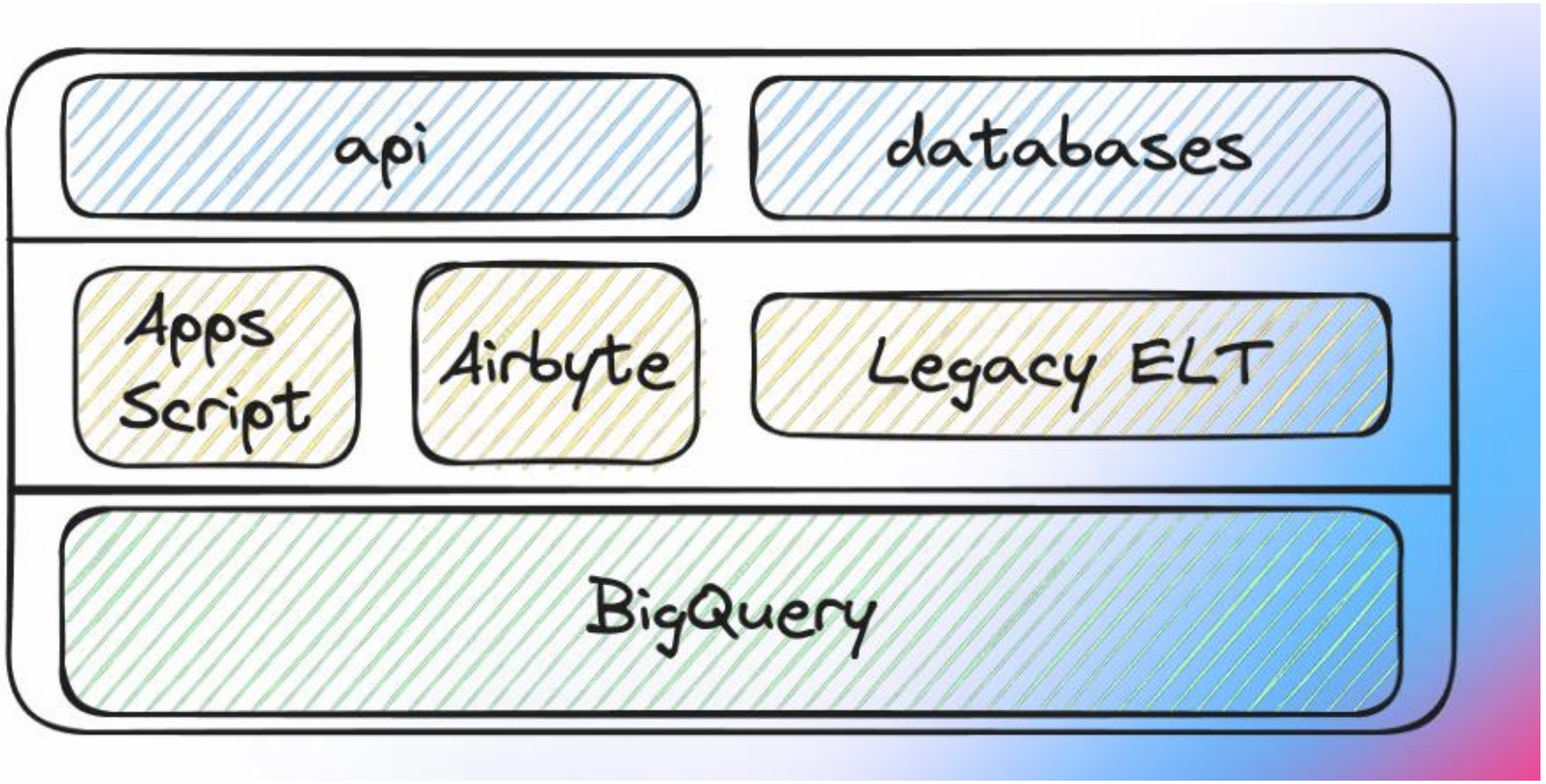
04

Игроки на рынке

Tbank(~2018)

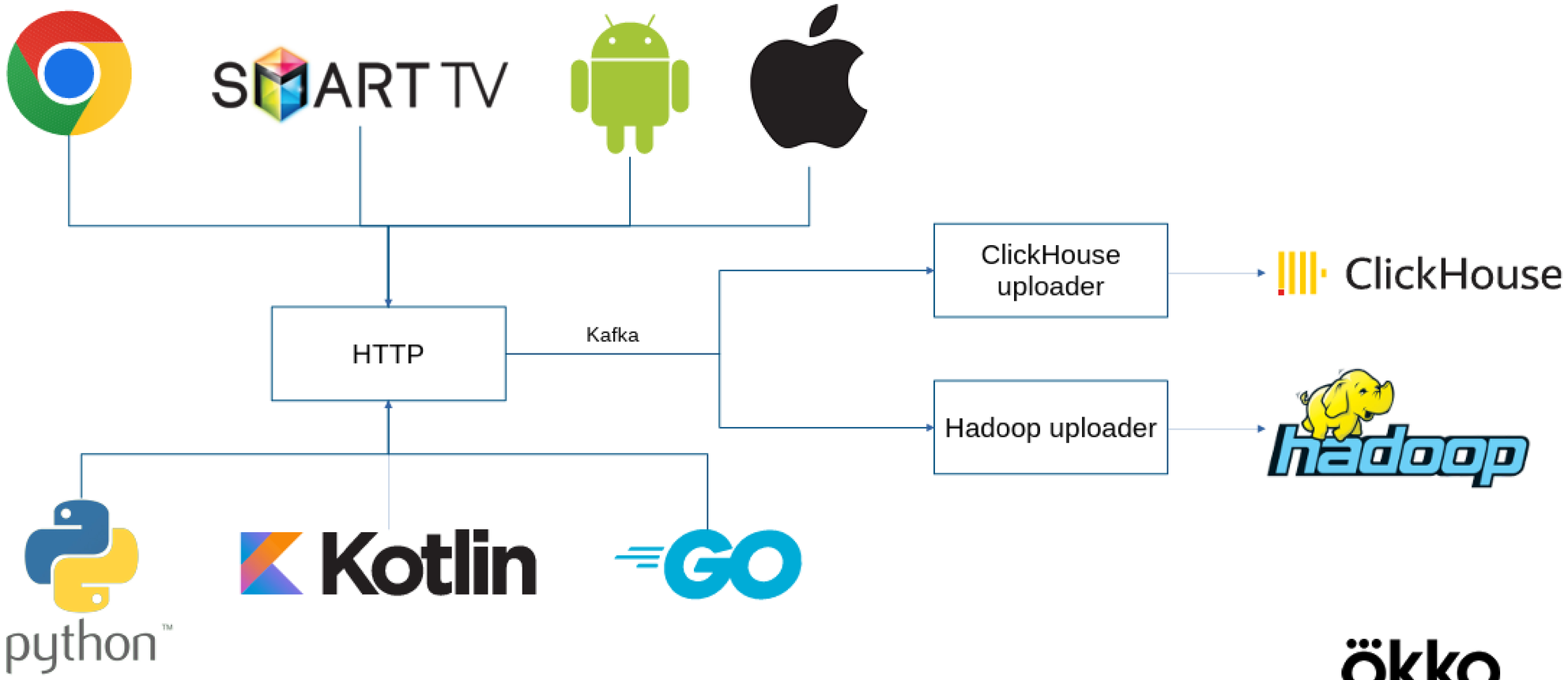

Общая архитектура DWH Tinkoff.ru
ETL/ELT
NAS (files)
oracle
oracle
oracle
ODS
DC1
DC2
9

Сберздоровье(~2022)

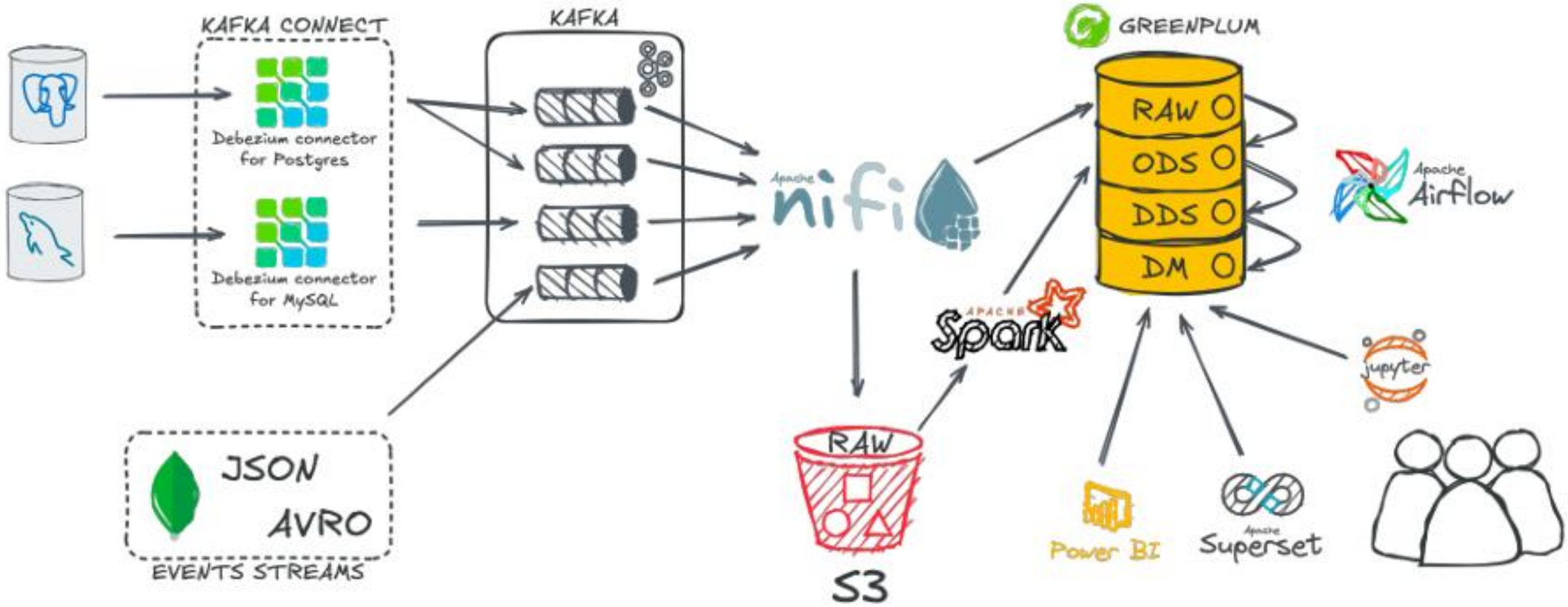


Okko(~2023)

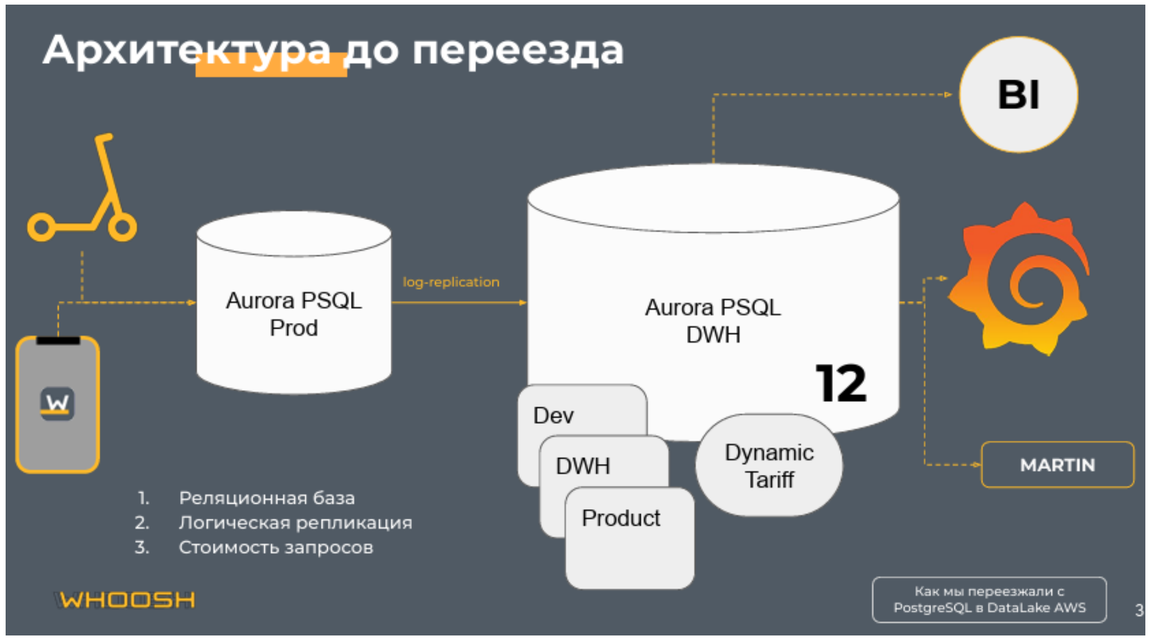
Архитектура

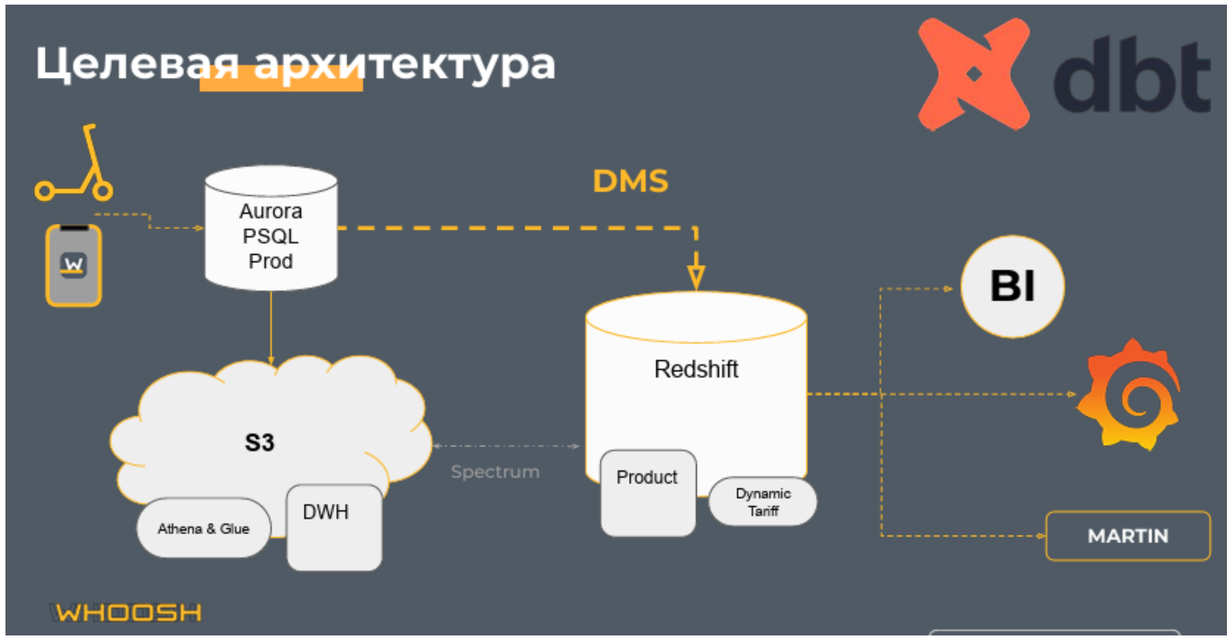


LemanaPro(~2022)

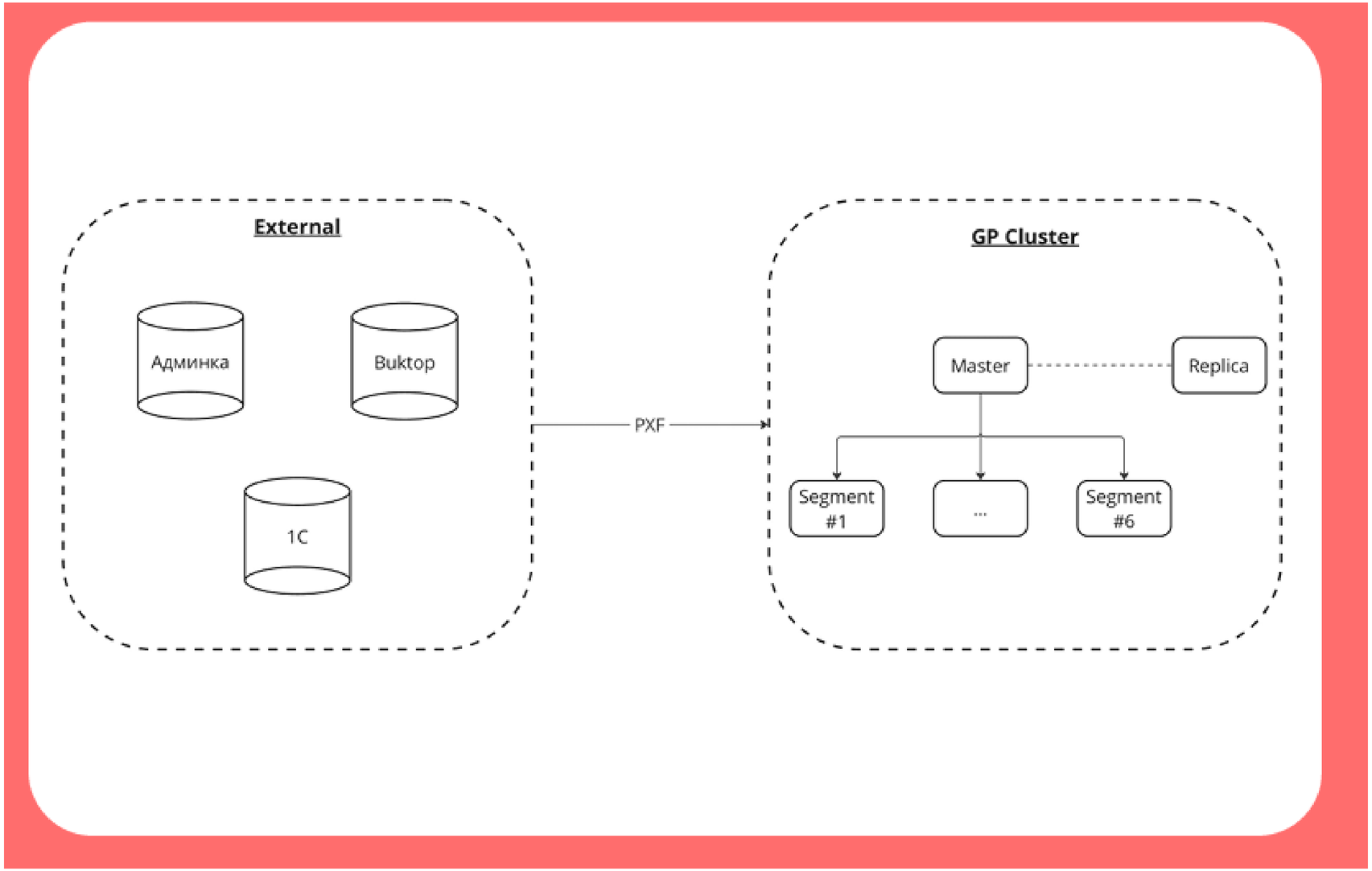


Whoosh(~2023)

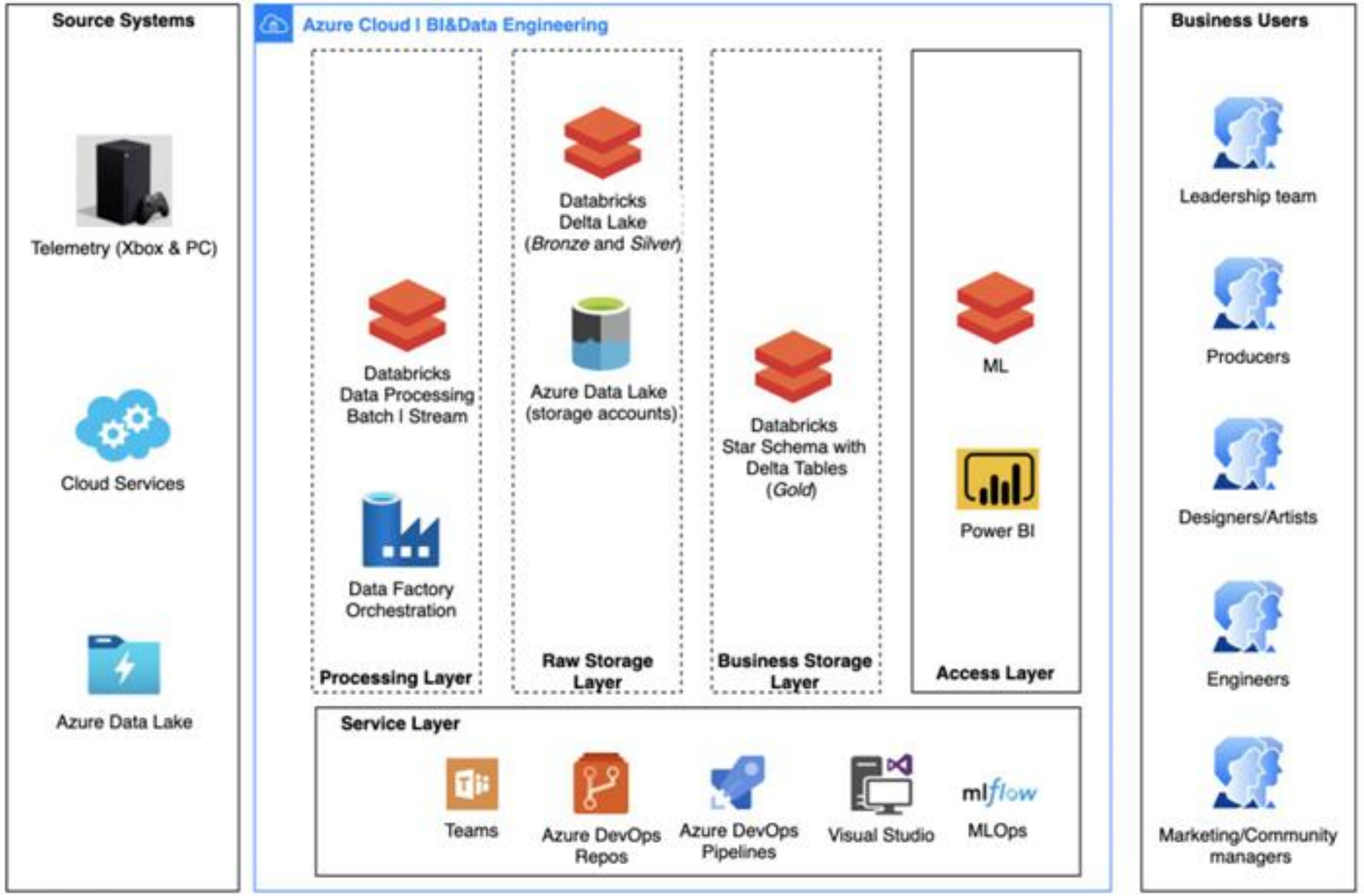

Архитектура до переезда
BI
Aurora PSQL Prod
log-replication
Aurora PSQL DWH
12
Dev
DWH
Product
Dynamic Tariff
MARTIN
1. Реляционная база
2. Логическая репликация
3. Стоимость запросов
WHOOSH
Как мы переезжали с PostgreSQL в DataLake AWS
3


Целевая архитектура
dbt
DMS
Aurora PSQL Prod
BI
Redshift
S3
Spectrum
Athena & Glue
DWH
Product
Dynamic Tariff
MARTIN
WHOOSH

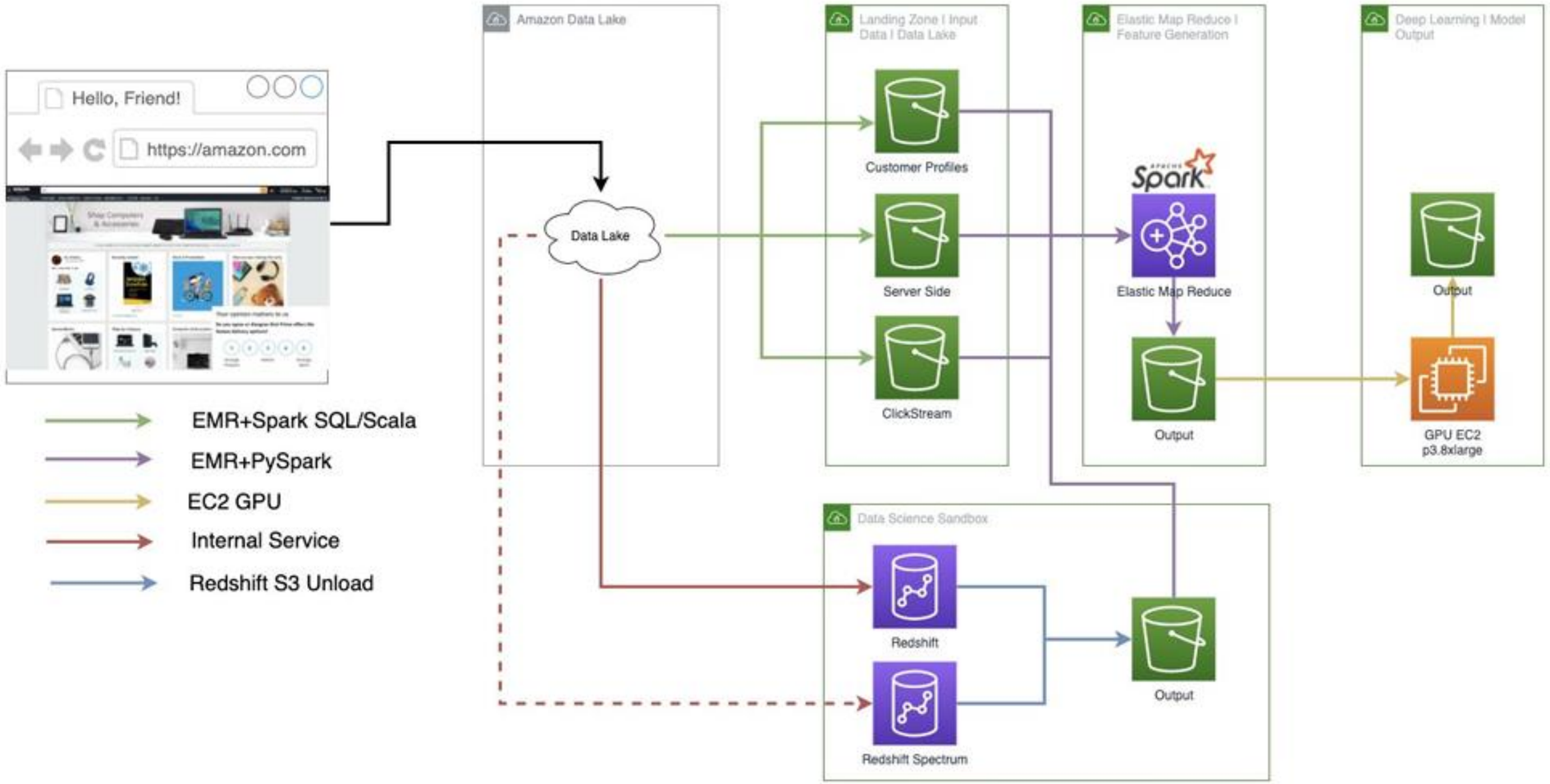
BestDoctor(~2022)



Xbox Microsoft(~2022)

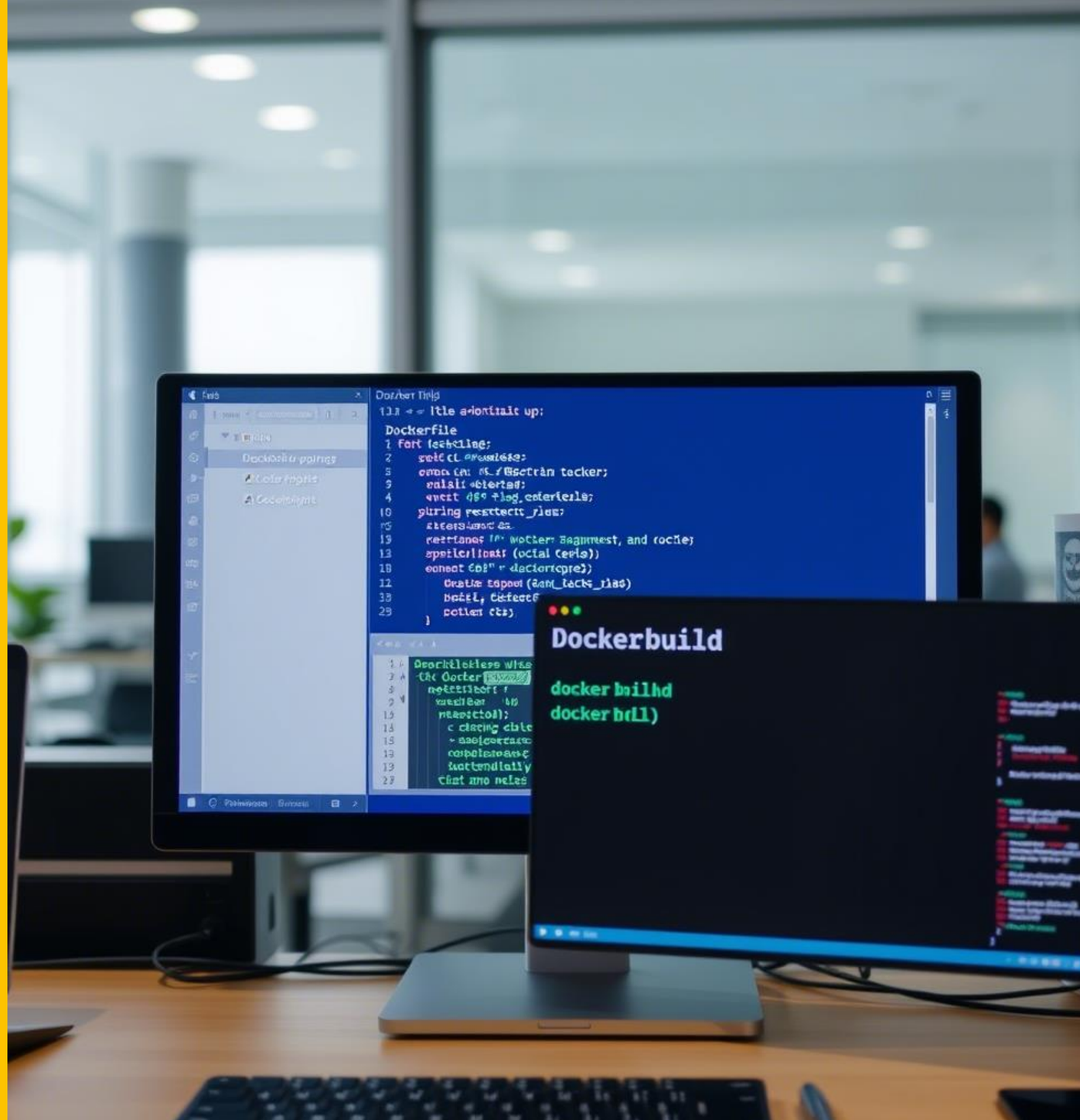


Amazon Retail(~2022)



05

Немного практики



Настройка

1. https://www.docker.com/products/docker-desktop/
2. https://github.com/zezOtik/bmstu--iu8-big-data-tools
3. https://dbeaver.io/