LENTIQ
a Bigstep company

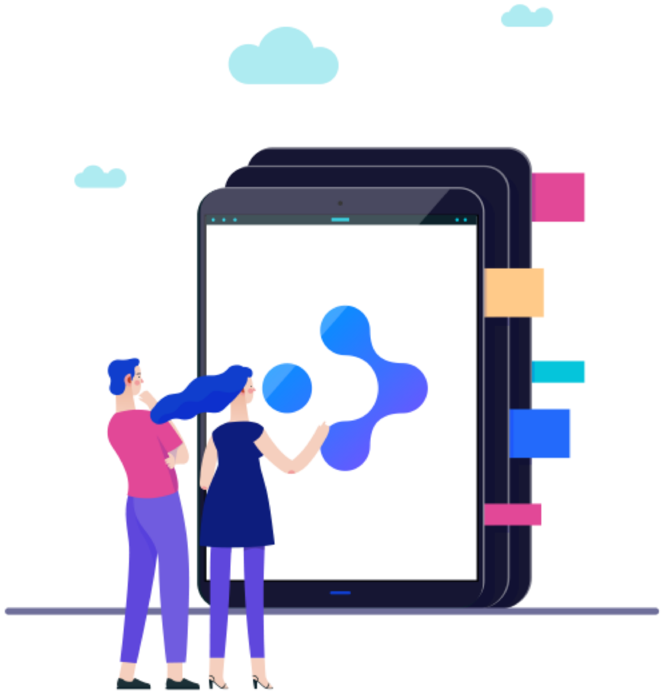
How to build the next generation data lake

Interconnected data pools
in a multi-cloud environment

Cristina Grosu – Product Manager @ Lentiq

Agenda

- Data lake design patterns
- The next generation data lake architecture powered by interconnected data pools
- Quick product walkthrough of Lentiq EdgeLake



About Lentiq

How we got here

Lentiq is an American company headquartered in Chicago, USA. It is a "spinoff" of Bigstep, a bare-metal cloud provider that helps companies run big data, machine learning, and analytics projects.

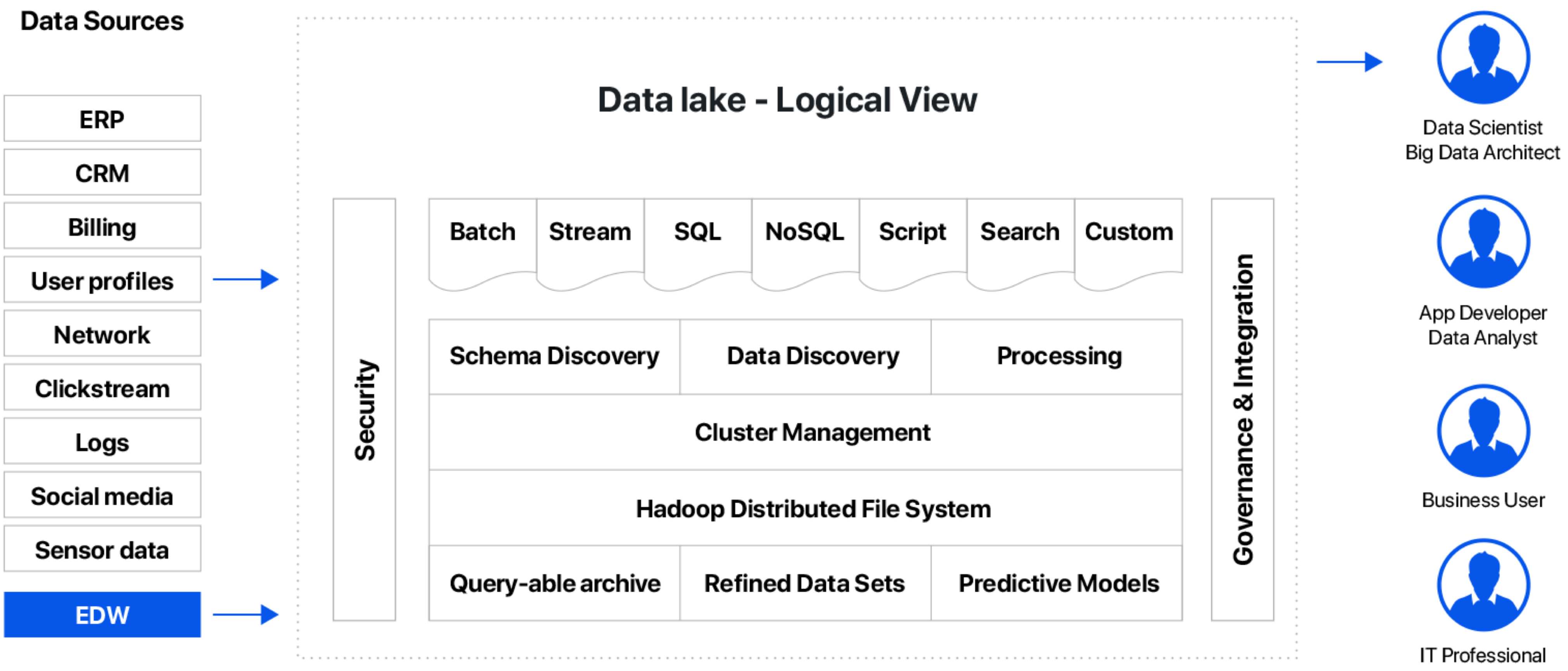
What we do

We focus on building data lakes that enable freedom and flexibility. We moved away from a centralized data repository to a fully distributed architecture that allows organizations to unify departments through data and knowledge sharing mechanisms.

What is a data lake?

An environment where data in multiple formats can be stored, accessed, processed, modelled, automated and visualized by a cross functional team in order to answer a wide array of business questions.

Data lake pattern based on Hadoop



Data lake pattern based on cloud-native services

aws

- Amazon EMR
- Amazon S3
- Amazon Kinesis
- Amazon Redshift
- Amazon DynamoDB
- Amazon RDS
- Amazon Lambda
- Amazon Athena
- Amazon Glue
- Amazon Quicksight

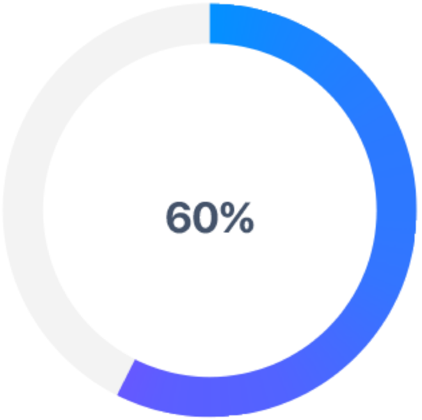
Google Cloud

- Google Dataproc
- Google Cloud Storage
- Google Dataflow
- Google Pub/Sub
- Google Cloud SQL
- Google BigQuery
- Google Datalab
- Google ML Engine
- Google Bigtable
- Google Spanner

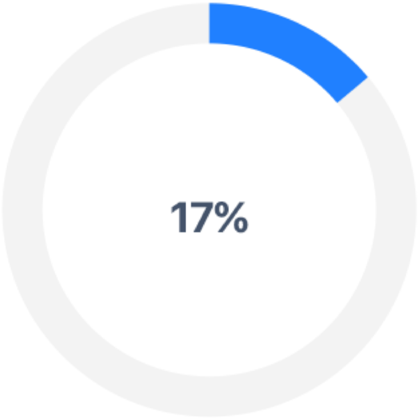
Azure

- Azure ADL Store
- Azure ADL Analytics
- Amazon HDInsight
- Azure Machine Learning
- Azure IoT Hub
- Azure SQL DW
- Azure Databricks
- Azure Data Factory
- Azure CosmosDB
- Azure PowerBI

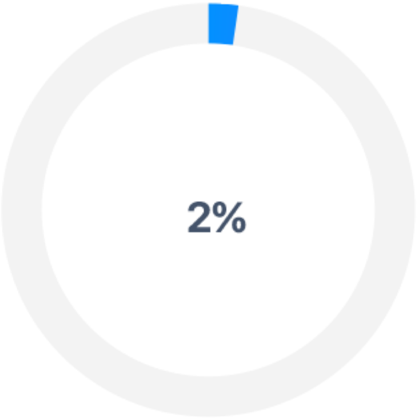
Current state of the data lakes



Data Lake projects failed to survive pilot phase



Deployments went to production



Achieved their initial goal

* According to Gartner and McKinsey research

Current data lakes problems

Over-centralized

All data projects must use the same technologies, schema model regardless of their organizational impact.

Overly-generalized

Current data lakes are built for the entire enterprise. This rigidity makes it harder to choose the right tools for a specific problem.

Complex

For all possible use cases, you will have Hadoop, key value stores, advanced data management and data lineage systems.

Expensive

The data lake implementation takes months, the project TCO is high and the team is complex (devops, big data engineeers, analysts).

Street talk

"I worked so much to standardize data before I put it in the central data lake, only to discover that I don't need it."

Data Scientist, Retail Company

"Tried to analyze some data in the central data lake, but customization required was implemented by central IT in a year. I wish I could have done it myself"

Lead of Analytics, Telecom Company

"They ask me to clean the data before I put it in, but I don't have the resources to do that here and I don't know yet if it is worth the effort"

Data Architect, Local Branch of Telecom Company

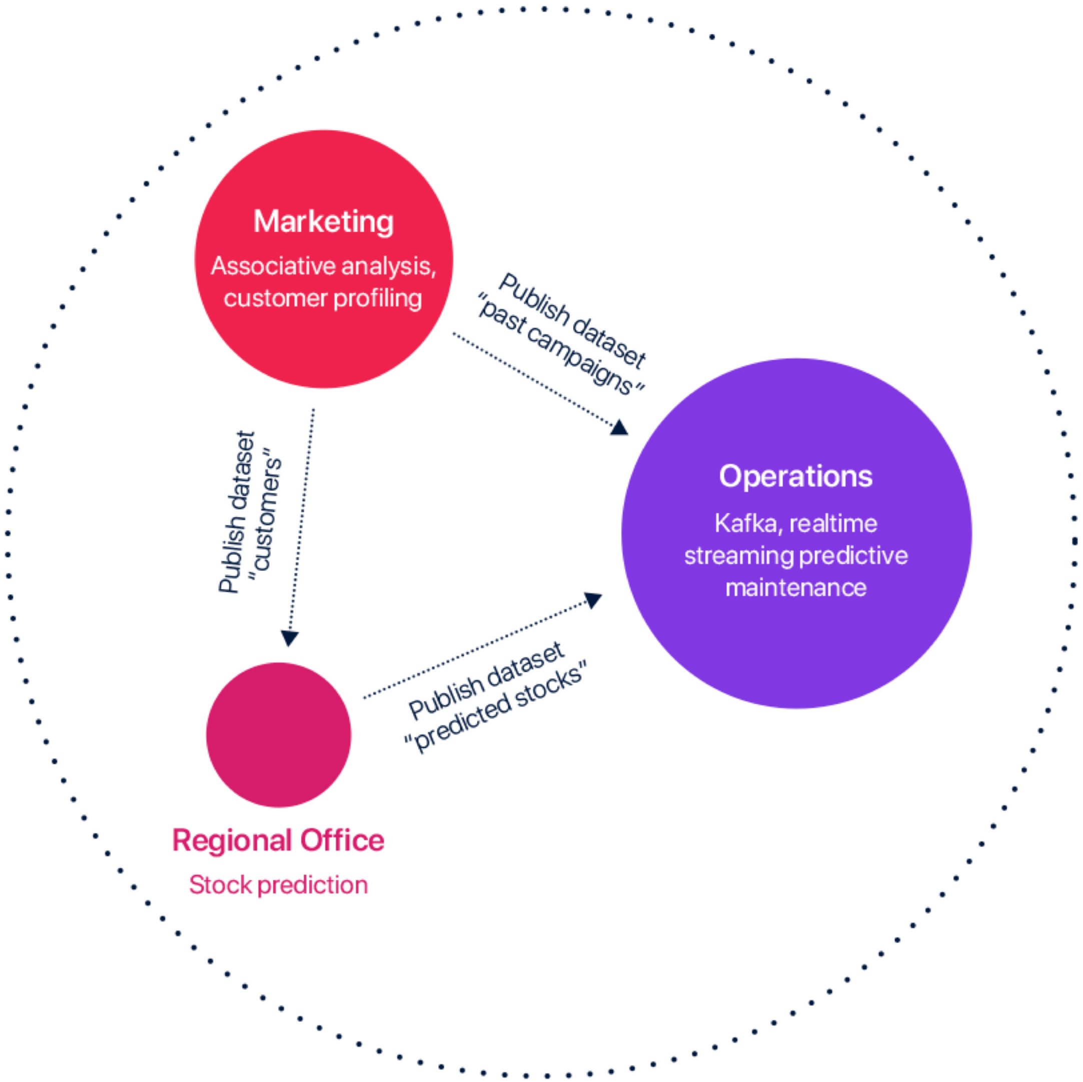
"Giving independence to local teams while maintaining standardization with current solutions seems impossible"

Director of Data Lake Platform, Banking Company

Balance flexibility and governance using data pools

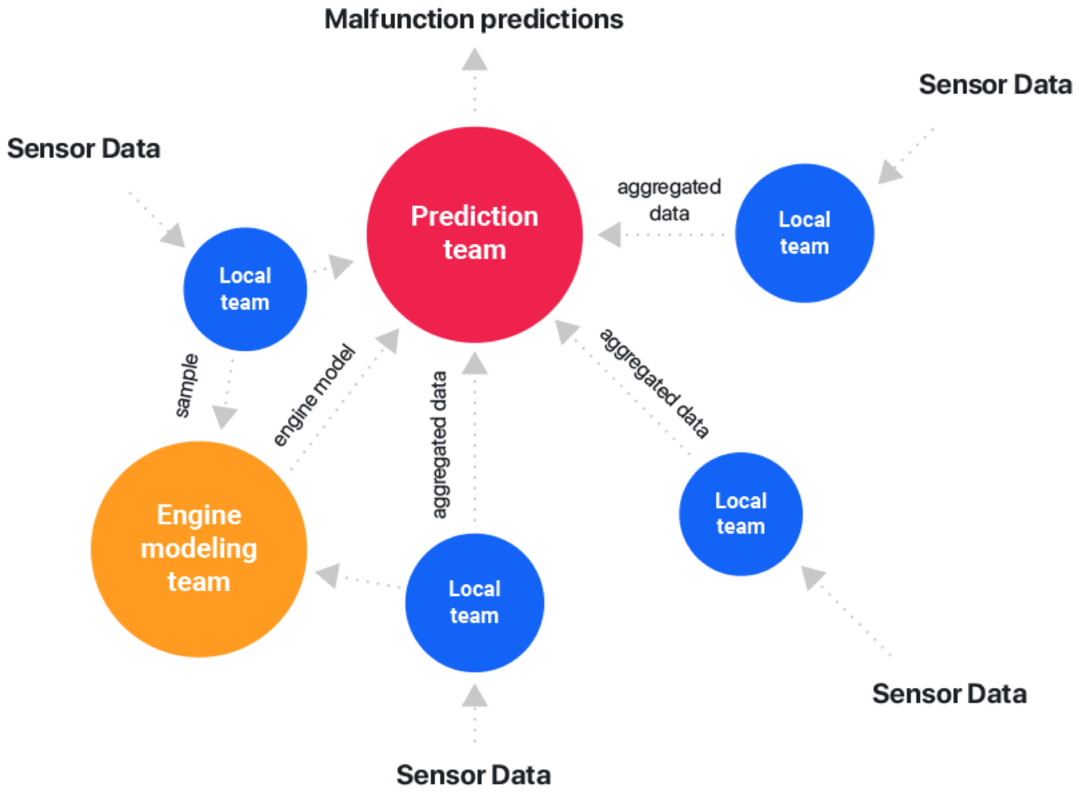
Transform the way you manage advanced analytics through our unique data pools architecture and a publish/subscribe data store:

- ✓ **Better data governance** via curated data sharing rather than dumping
- ✓ **More independence** for business units to run their own stack and solve their own problems
- ✓ **More affordable** using pay-per-use services
- ✓ **Increased adoption** through intuitive, single pane of glass user experience
- ✓ **Multi-cloud**: different data pools can run on different cloud vendors


Marketing
Associative analysis, customer profiling
Publish dataset "past campaigns"
Publish dataset "customers"
Operations
Kafka, realtime streaming predictive maintenance
Publish dataset "predicted stocks"
Regional Office
Stock prediction

Predictive maintenance use case

Predict malfunctions based on real time sensor data and component models.



Introducing EdgeLake

A flexible, decentralized data lake service, spanning multiple clouds and regions, enabling independent development while fostering collaboration.

Unified management

Unified management regardless of the underlying infrastructure provider

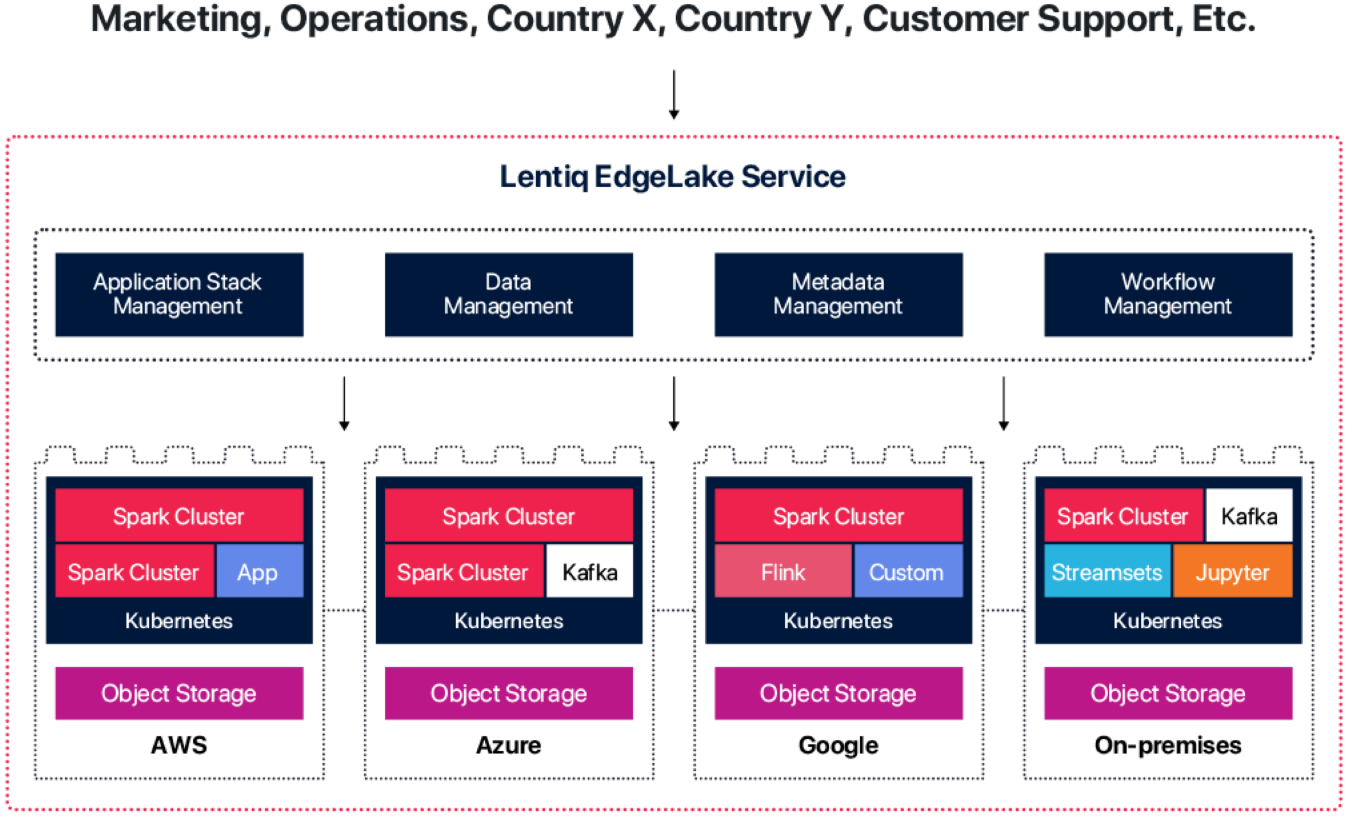
Faster innovation through self-service

The right stack for the your team and use case

Lower maintenance costs

No ops or specialized skills

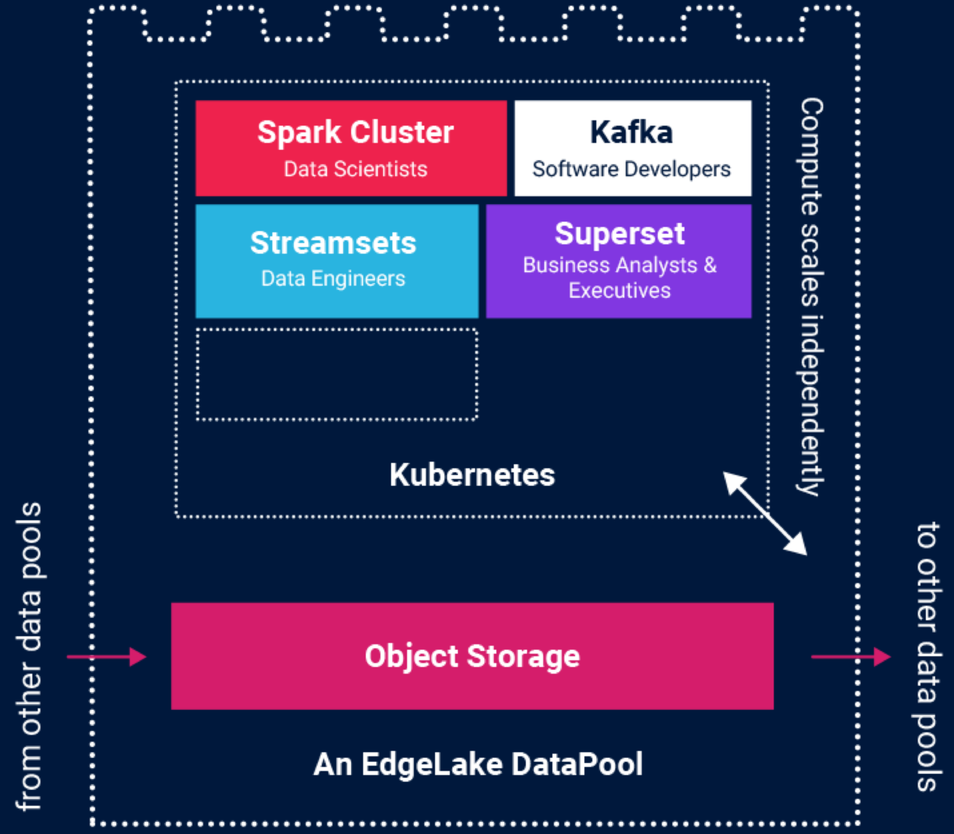
Lentiq's approach: Interconnected Data Pools



What is a "data pool"?

A data pool is a micro-data lake. It provides everything a data scientist or data engineer needs: data management capabilities, notebook environments, Apache Spark cluster management, and others.

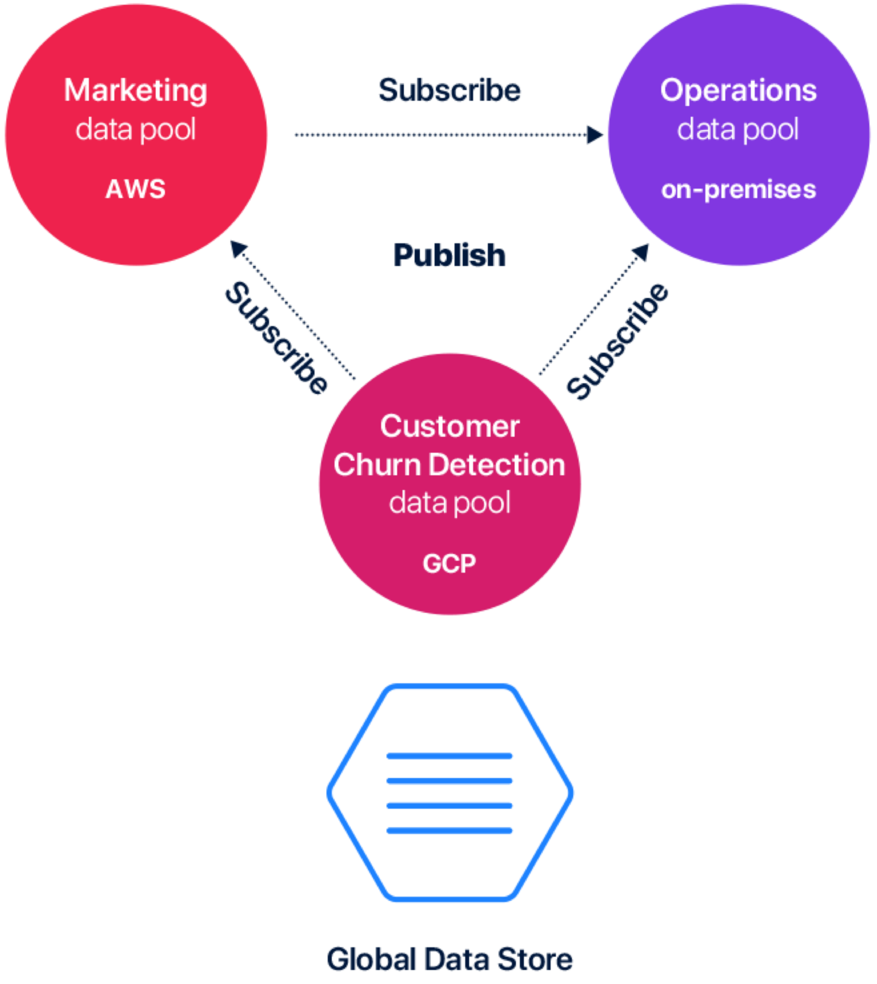
- ✓ **Independent**: each data pool has its own budget and resources
- ✓ **Flexible**: in a data pool, you can have the best tools needed for each specific business use case
- ✓ **At the edge**: closer to the data source and the data team that uses it



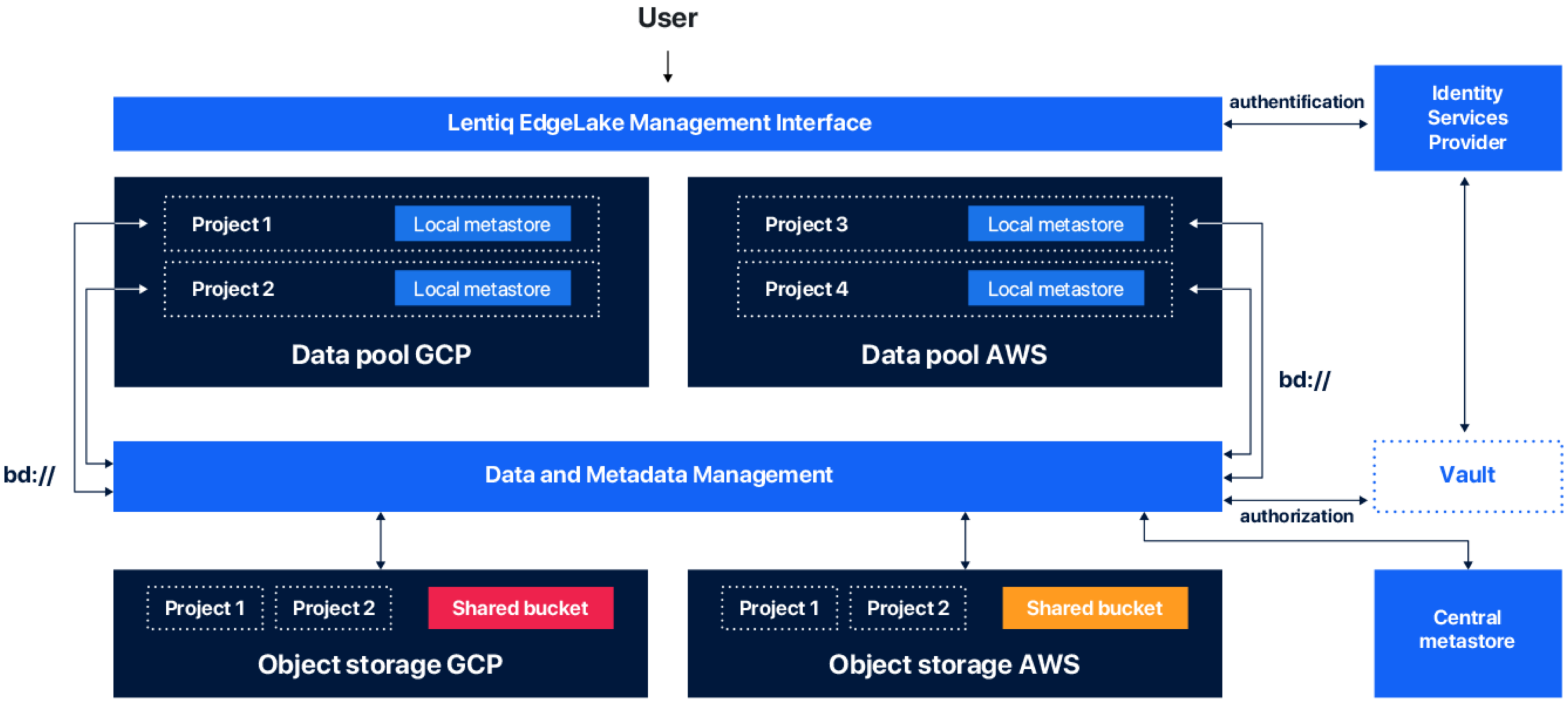
Collaborate through the global data store

Our publish-subscribe data concept allows maximum flexibility at project level and enforces data documentation when sharing data to the rest of the organization.

- ✓ **Project-tailored data**
  At project level, use data in the format needed for maximum insights, without worrying of standardization and governance.
- ✓ **Curate and document datasets**
  Annotate tables, columns and files with descriptions, comments and tags and increase explainability of data and adoption.
- ✓ **Publish curated datasets**
  Make your data available to the rest of the organization and inspire experimentation in other teams.



Data and metadata management



It's demo time!

Lentiq EdgeLake at a glance

Unified management

Regardless of the underlying infrastructure provider

Faster innovation through self-service

Choose the right stack for your team and business use case.

Lower maintenance costs

No ops or specialized skills

Data and Knowledge Sharing

- ✓ Share curated datasets with the rest of the organization
- ✓ Share curated notebooks with the rest of the organization
- ✓ Connect with other data teams

Open Source Application Management

- ✓ Code in Python
- ✓ Jupyter Notebooks as a Service
- ✓ Apache Spark as a Service
- ✓ Apache Kafka as a Service
- ✓ PostgreSQL as a Service
- ✓ Top Python libraries: Pandas, Ray Numpy, Dask, Seaborn, XGBoost Matplotlib, Scikit-learn, Spark ML
- ✓ Provision clusters and scale them as needed

Data Science and AI at Scale

- ✓ Data science at scale through Dask, Spark and Ray
- ✓ Run Spark jobs on independent clusters
- ✓ Model management and deployment
- ✓ Provision use case specific projects with their own budget are resources

Data and Metadata Management

- ✓ Annotate files and tables before sharing
- ✓ Create tables from files through Spark and explore them in the table browser
- ✓ Document table columns before sharing and improve data set explainability and adoption

Q&A

Thank you