A modern adatplatform receptje

Arató Bence
arato@biconsulting.hu

Arató Bence

A BI Consulting igazgatója, 20 éves szakmai múlttal az adattárházak, BI és analitika területén

- **Adatstratégiai tanácsadás**

  Adat- és BI stratégia kialakítása, eszközválasztási tanácsadás, pályáztatás, minőségbiztosítás, mentoring

- **Oktatás**

  A BI Akadémia vezető oktatójaként szakterületei közé a DW/BI architektúrák, Big Data eszközök, adatmodellezés és adatvizualizáció tartozik

- **Publikációk**

  Az Adat.blog és a Dataviz.hu szerkesztője

- **Rendezvények**

  A Budapest Data Fórum és Budapest BI Fórum konferenciák vezető szervezője.

- **Meetupok**

  A Big Data, a BI és a PyData meetupok szervezője

BI Consulting

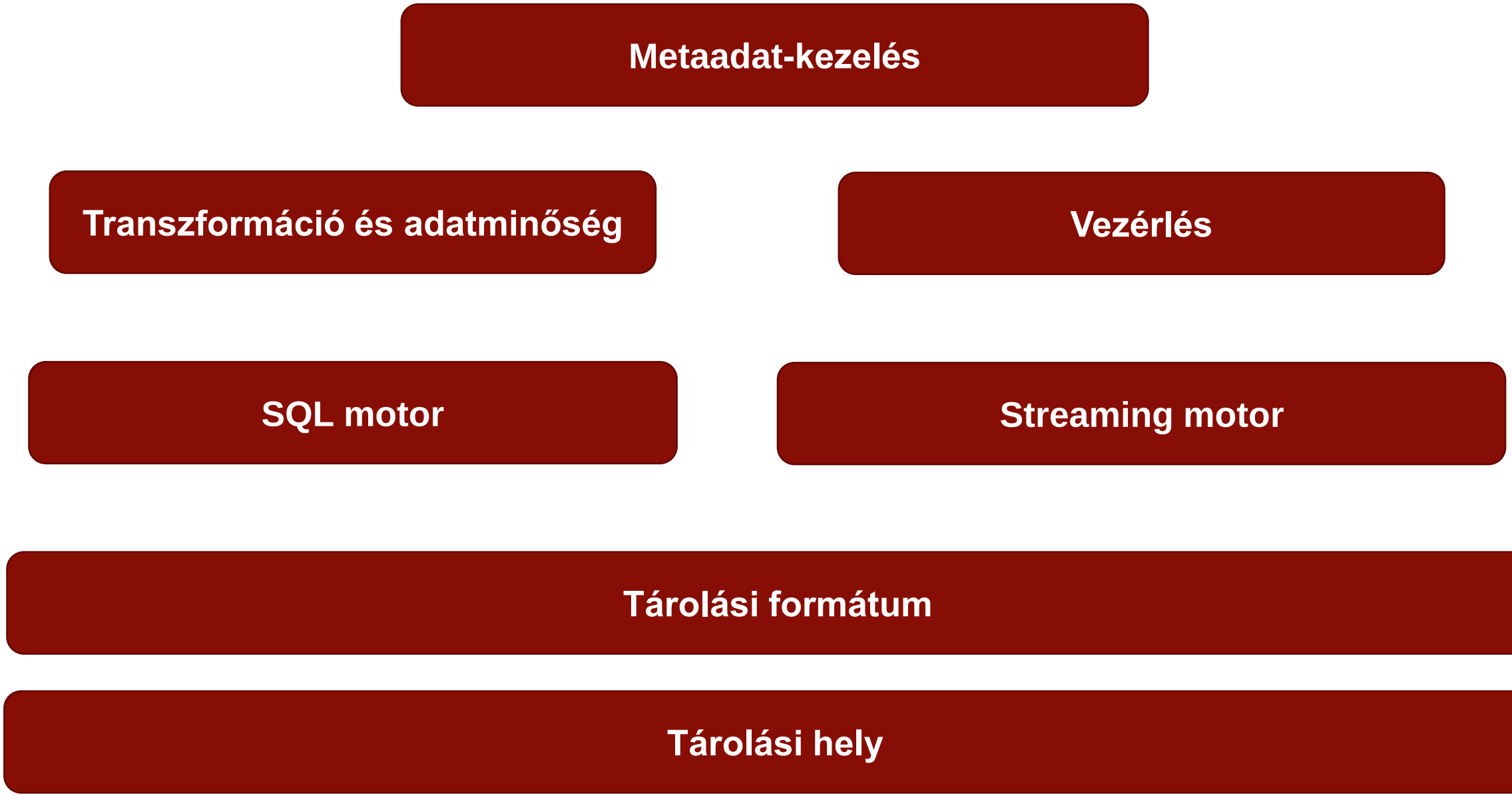
Modern adatplatformok

BI Consulting

Milyen is a modern adatplatform?

- Legyen hatékony, ezért fontos a Storage / Compute szétválasztása, ahol az adatok tárolása és feldolgozása két külön réteg
- Legyen univerzális, azaz fusson on-premise vagy bármelyik nagyobb felhőben
- Legyen nyílt forráskódú a legtöbb komponens

Az adatplatform receptje

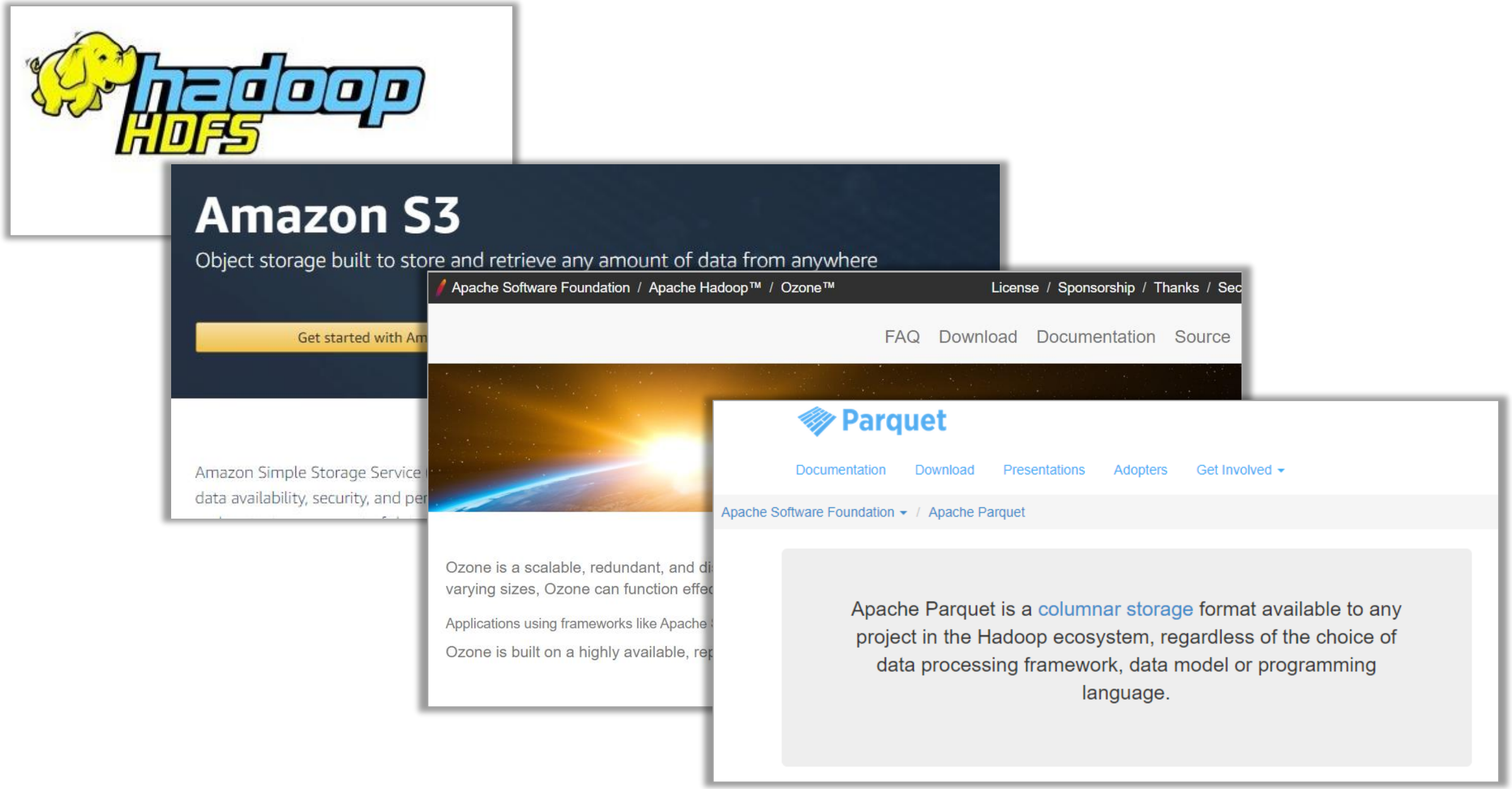


Tárolás

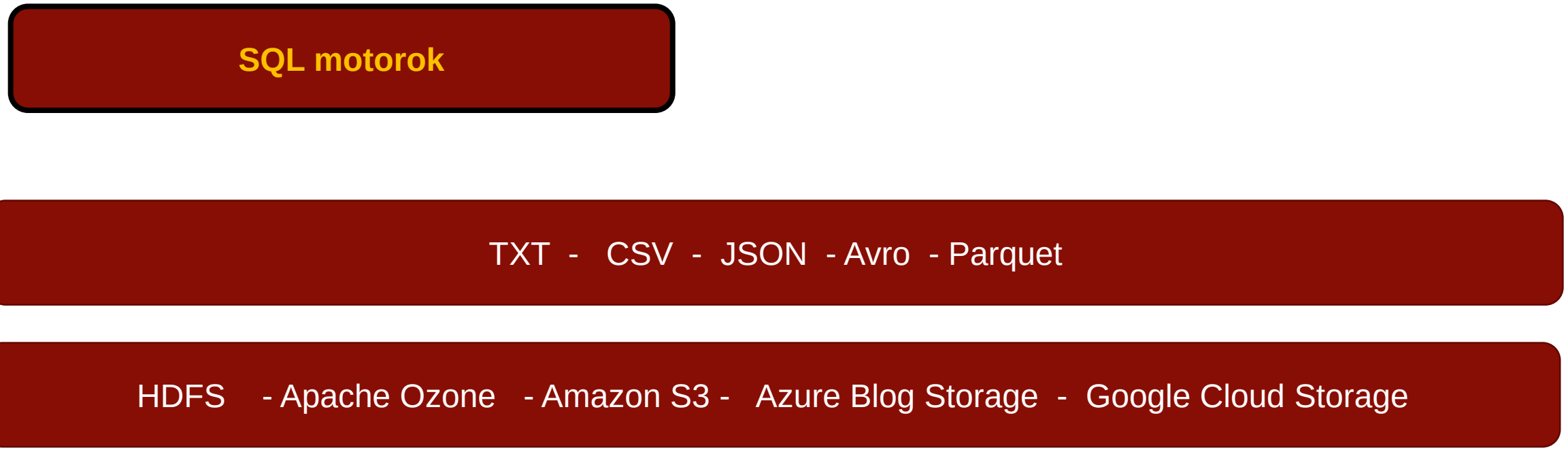
Tárolási formátum

Tárolási hely

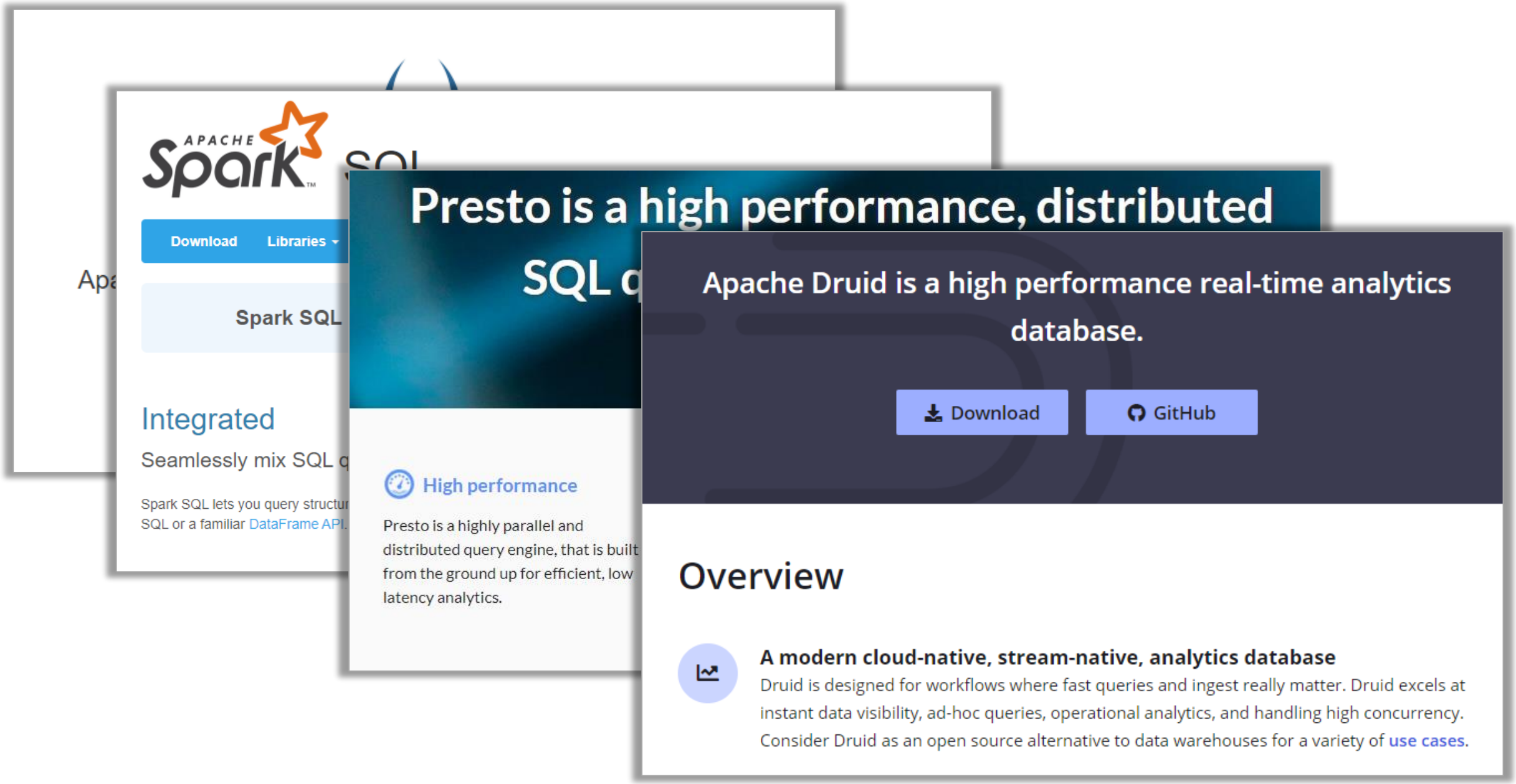
Tárolás


hadoop
HDFS
Amazon S3
Object storage built to store and retrieve any amount of data from anywhere
Get started with Am
Amazon Simple Storage Service
data availability, security, and per
Apache Software Foundation / Apache Hadoop™ / Ozone™
License / Sponsorship / Thanks / Sec
FAQ Download Documentation Source
Ozone is a scalable, redundant, and di
varying sizes, Ozone can function effec
Applications using frameworks like Apache
Ozone is built on a highly available, rep
Parquet
Documentation Download Presentations Adopters Get Involved
Apache Software Foundation / Apache Parquet
Apache Parquet is a columnar storage format available to any project in the Hadoop ecosystem, regardless of the choice of data processing framework, data model or programming language.

SQL motorok


SQL motorok
TXT - CSV - JSON - Avro - Parquet
HDFS - Apache Ozone - Amazon S3 - Azure Blog Storage - Google Cloud Storage

SQL motorok


APACHE Spark
Download
Libraries
Spark SQL
Integrated
Seamlessly mix SQL
Spark SQL lets you query structur
SQL or a familiar DataFrame API.
Presto is a high performance, distributed
SQL q
High performance
Presto is a highly parallel and distributed query engine, that is built from the ground up for efficient, low latency analytics.
Apache Druid is a high performance real-time analytics database.
Download
GitHub
Overview
A modern cloud-native, stream-native, analytics database
Druid is designed for workflows where fast queries and ingest really matter. Druid excels at instant data visibility, ad-hoc queries, operational analytics, and handling high concurrency. Consider Druid as an open source alternative to data warehouses for a variety of use cases.

Streaming motorok

Impala/Hive - Spark - Presto - Druid

Streaming motorok

TXT - CSV - JSON - Avro - Parquet

HDFS - Apache Ozone - Amazon S3 - Azure Blog Storage - Google Cloud Storage

Streaming motorok

APACHE KAFKA

*More than **80% of all Fortune 100 companies** trust, and use Kafka.*

Apache Kafka is an open
thousands of companies fo
data integr

10 OUT OF 10

MANUFACTURING

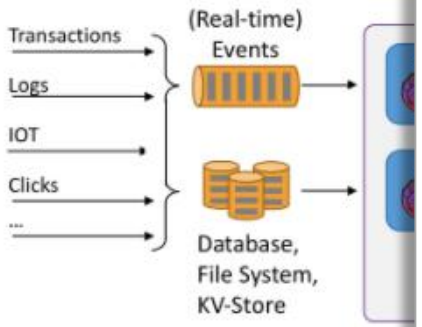
PULSAR

Apache Pulsar is a cloud-native, distributed messaging and streaming platform originally created at Yahoo! and now a top-level Ap

Pulsar Functions

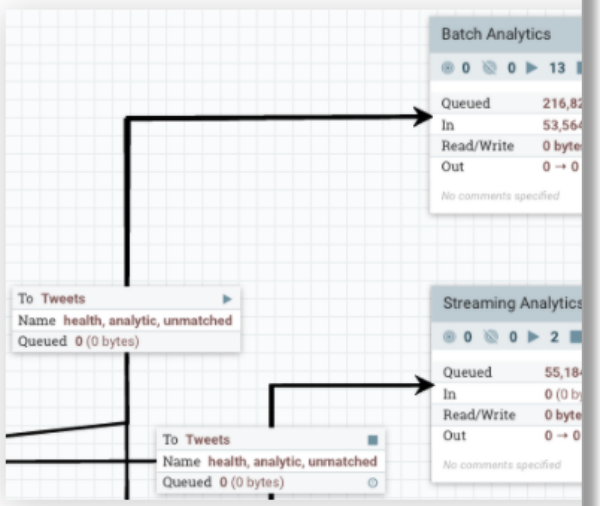
Easy to deploy, lightweight compu
process, developer-friendly APIs, no
to run your own stream processir
engine.

Apache Flink® — Stateful Computations over Data Streams



APACHE nifi

An easy to use, powerful, and reliable system to process and distribute data.



Transzformáció és adatminőség

Transzformáció és adatminőség

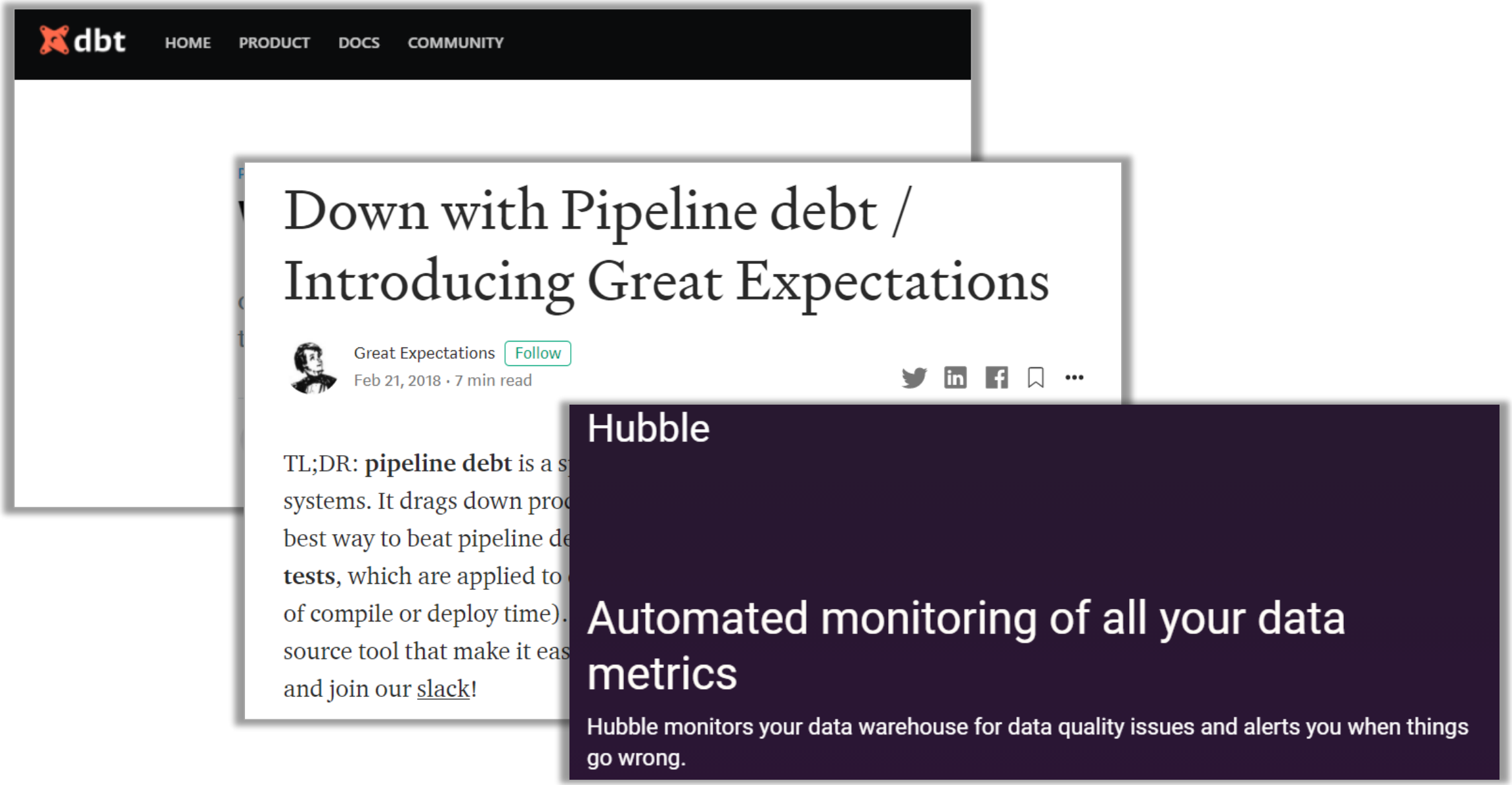
Impala/Hive - Spark - Presto - Druid

Kafka - Pulsar - Flink - Nifi - Beam - Spark

TXT - CSV - JSON - Avro - Parquet

HDFS - Apache Ozone - Amazon S3 - Azure Blog Storage - Google Cloud Storage

Transzformáció és adatminőség



Vezérlés

dbt - Great Expectations - Hubble

Vezérlés

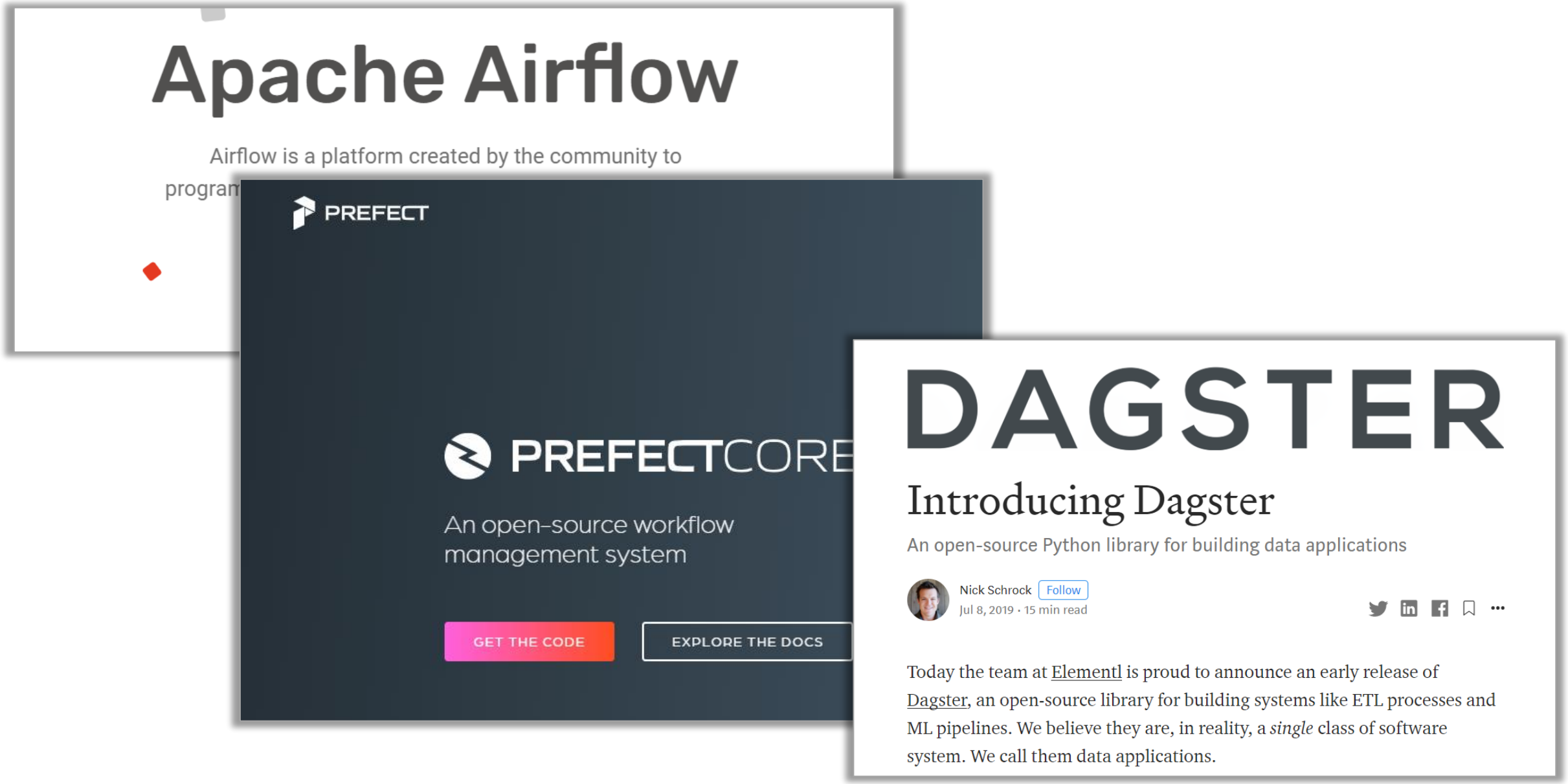
Impala/Hive - Spark - Presto - Druid

Kafka - Pulsar - Flink - Nifi - Beam - Spark

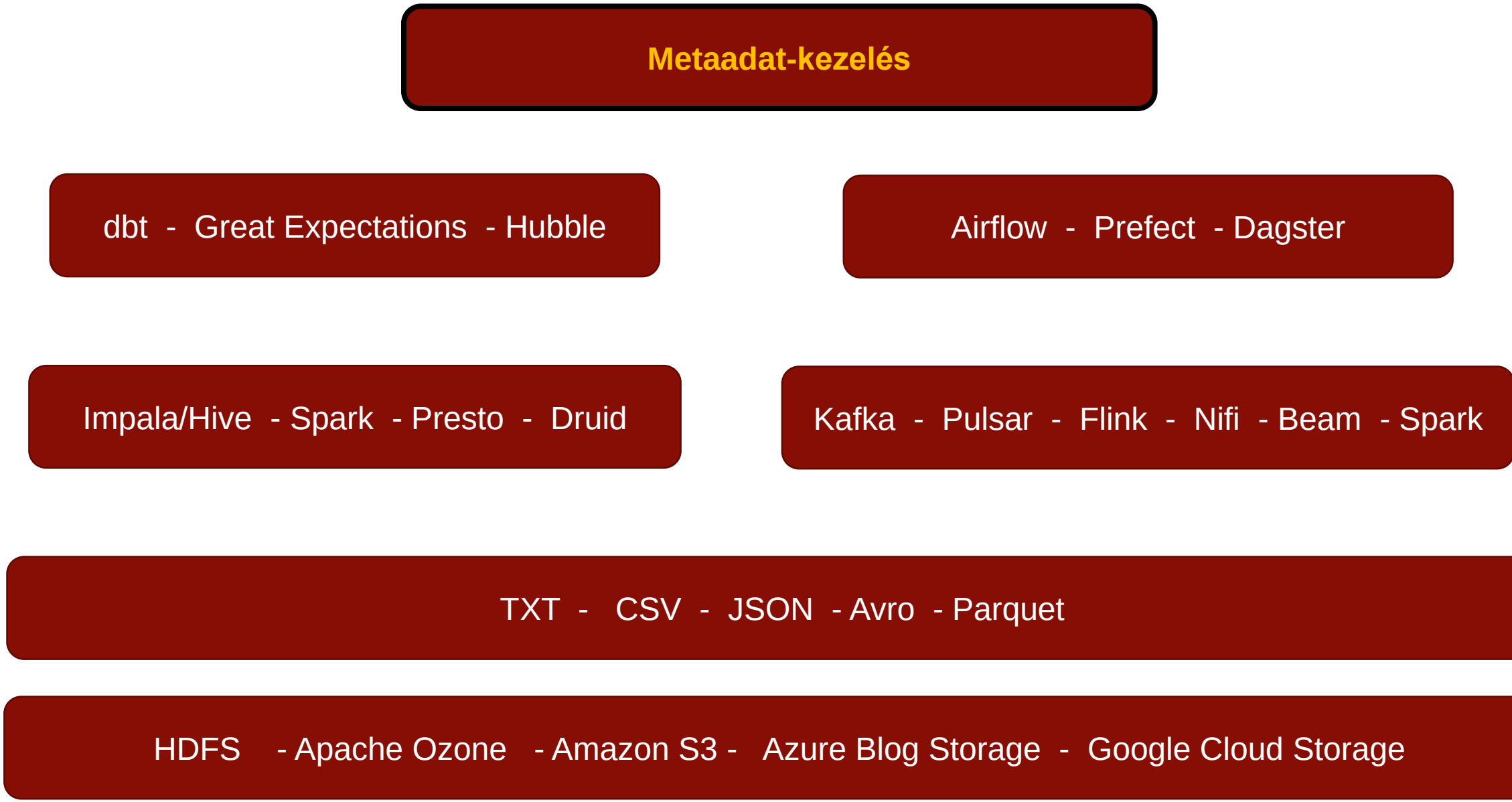
TXT - CSV - JSON - Avro - Parquet

HDFS - Apache Ozone - Amazon S3 - Azure Blog Storage - Google Cloud Storage

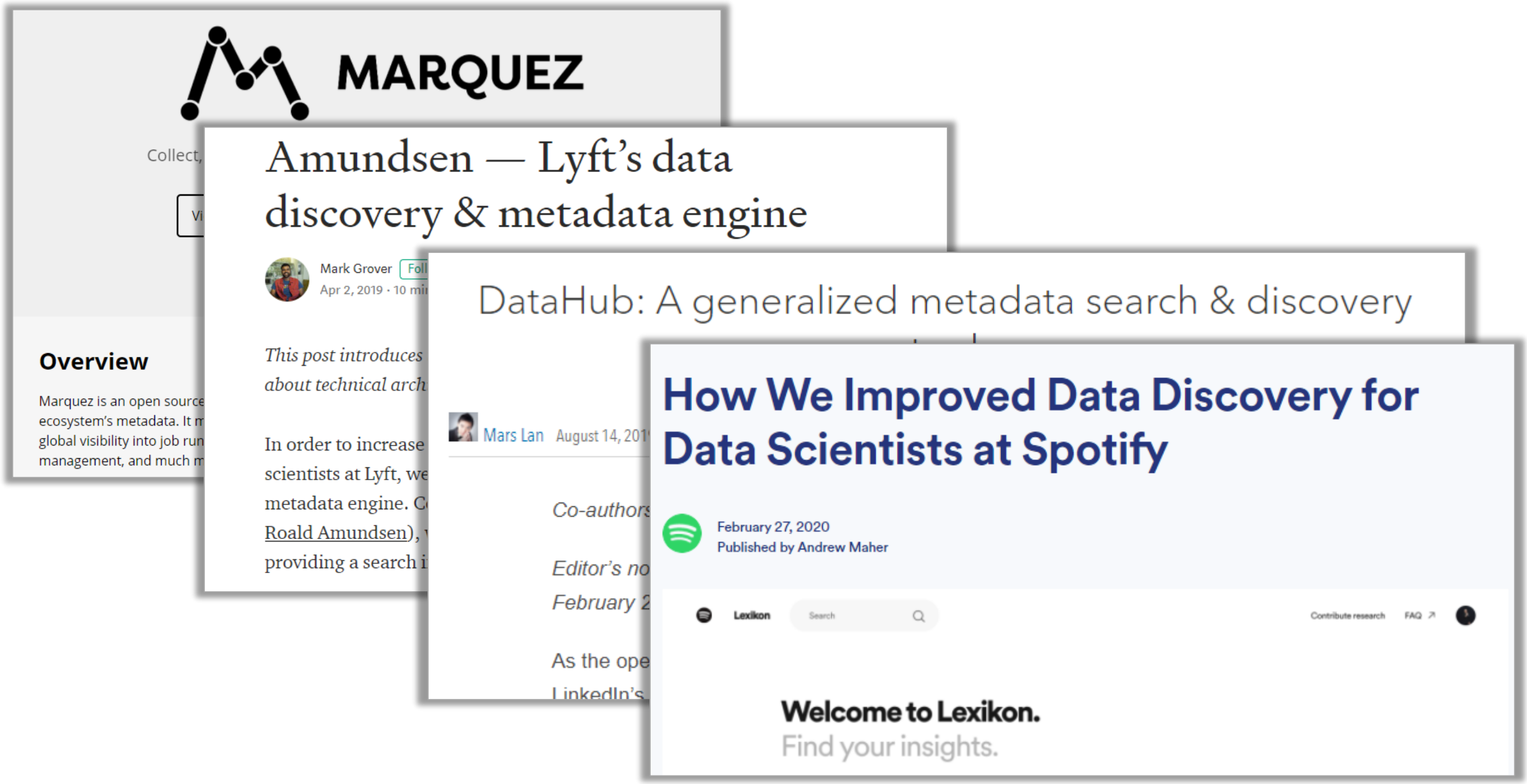
Vezérlés


Apache Airflow
Airflow is a platform created by the community to
PREFECT
PREFECTCORE
An open-source workflow
management system
GET THE CODE
EXPLORE THE DOCS
DAGSTER
Introducing Dagster
An open-source Python library for building data applications
Nick Schrock
Follow
Jul 8, 2019 · 15 min read
Today the team at Elementl is proud to announce an early release of
Dagster, an open-source library for building systems like ETL processes and
ML pipelines. We believe they are, in reality, a single class of software
system. We call them data applications.

Metaadat-kezelés



Metaadat-kezelés



Az adatplatform receptje

Amundsen - Marquez - Datahub - Lexikon

dbt - Great Expectations - Hubble

Airflow - Prefect - Dagster

Impala/Hive - Spark - Presto - Druid

Kafka - Pulsar - Flink - Nifi - Beam - Spark

TXT - CSV - JSON - Avro - Parquet

HDFS - Apache Ozone - Amazon S3 - Azure Blog Storage - Google Cloud Storage

A sárgával jelölt témákról lesz külön előadás

Amundsen - Marquez - Datahub - Lexikon

dbt - Great Expectations - Hubble

Airflow - Prefect - Dagster

Impala/Hive - Spark - Presto - Druid

Kafka - Pulsar - Flink - Nifi - Beam - Spark

TXT - CSV - JSON - Avro - Parquet

HDFS - Apache Ozone - Amazon S3 - Azure Blog Storage - Google Cloud Storage

Thank You

BI Consulting