

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



The slide features a dark blue background with a stylized world map composed of light blue dots. Various colored circles (red, orange, yellow, purple) are scattered across the map and background. A prominent red diagonal line runs from the top right towards the bottom right. A dark red horizontal band is positioned across the middle of the slide, containing the title text in white. The bottom left corner contains the BI Consulting logo, which consists of three horizontal red bars inside a white circle, followed by the text 'BI Consulting' in white.

Modern adatplatformok

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

Metaadat-kezelés

Transzformáció és adatminőség

Vezérlés

SQL motor

Streaming motor

Tárolási formátum

Tárolási hely

Tárolási formátum

Tárolási hely

Tárolás



Amazon S3

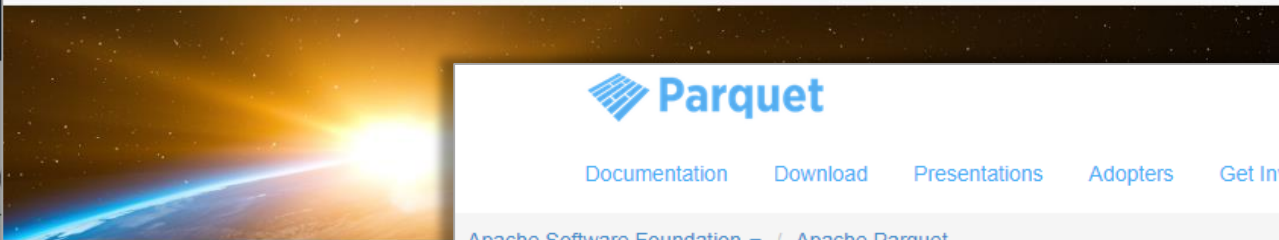
Object storage built to store and retrieve any amount of data from anywhere

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

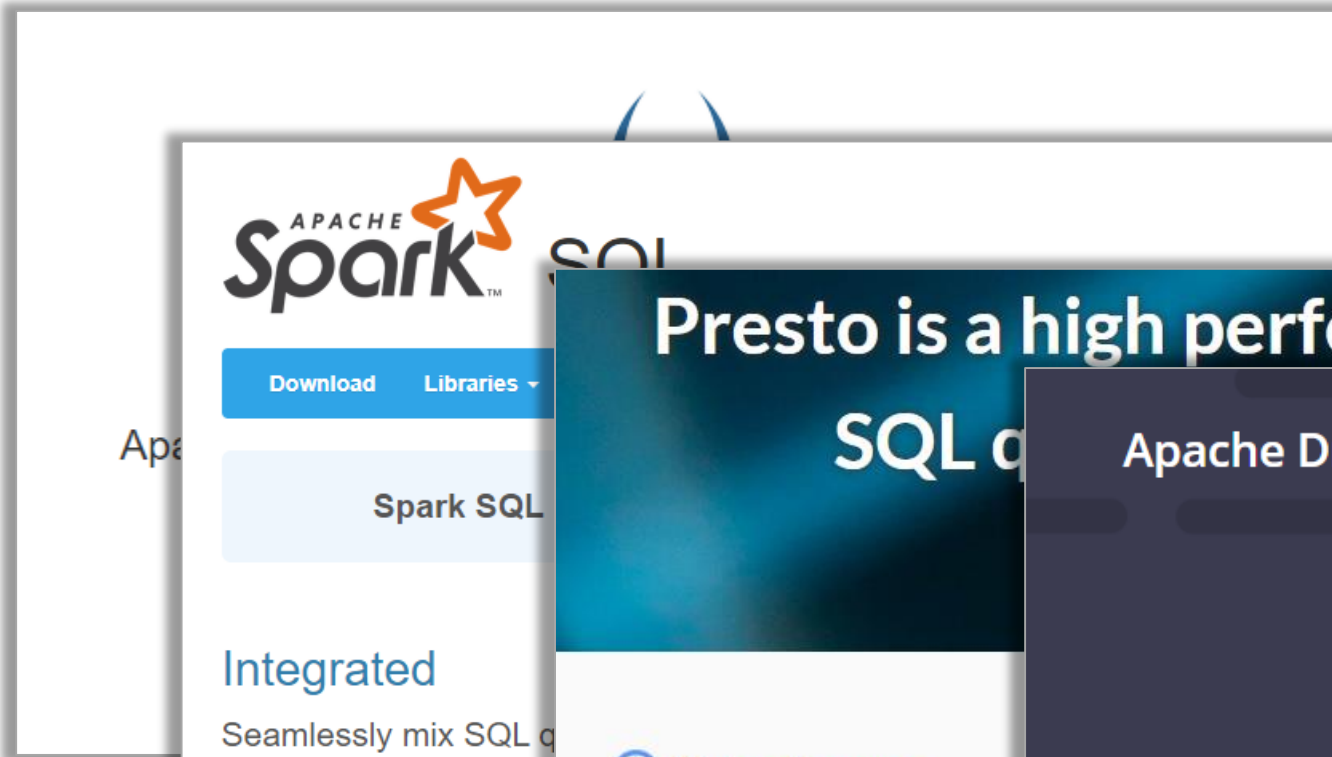
Apache Parquet is a **columnar storage** format available to any project in the Hadoop ecosystem, regardless of the choice of data processing framework, data model or programming language.

SQL motorok

TXT - CSV - JSON - Avro - Parquet

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

SQL motorok



APACHE **Spark**™ SQL

Download Libraries ▾


Spark SQL

Integrated

Seamlessly mix SQL and Spark

Spark SQL lets you query structured data using SQL or a familiar [DataFrame API](#).

Presto is a high performance, distributed SQL engine


 **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 Blob 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 source distributed streaming platform that powers thousands of companies for data integration and analytics.



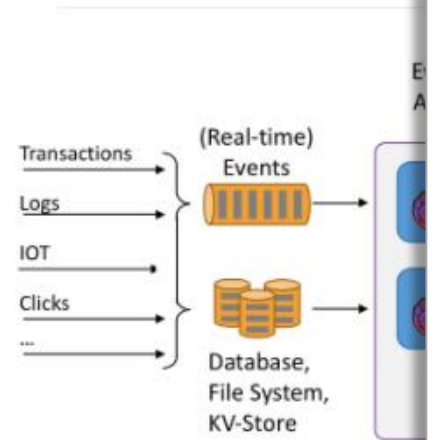
10 OUT OF 10
MANUFACTURING



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

Pulsar Functions

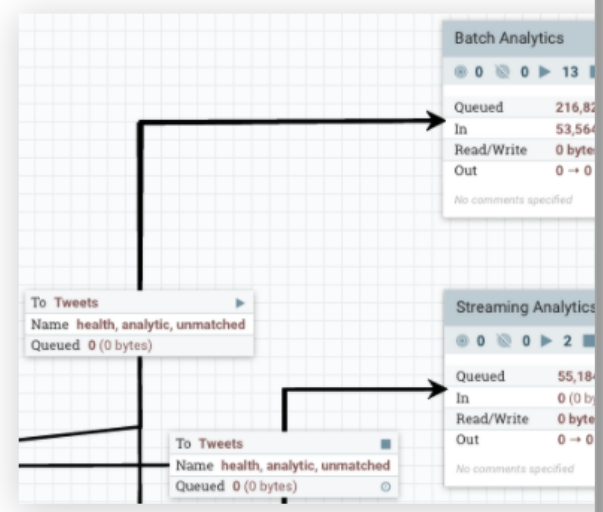
Easy to deploy, lightweight component, process, developer-friendly APIs, no need to run your own stream processing engine.



Apache Flink® — Stateful Computations over Data Streams



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



Transzformáció és adatminőség

Transzformáció és adatminőség

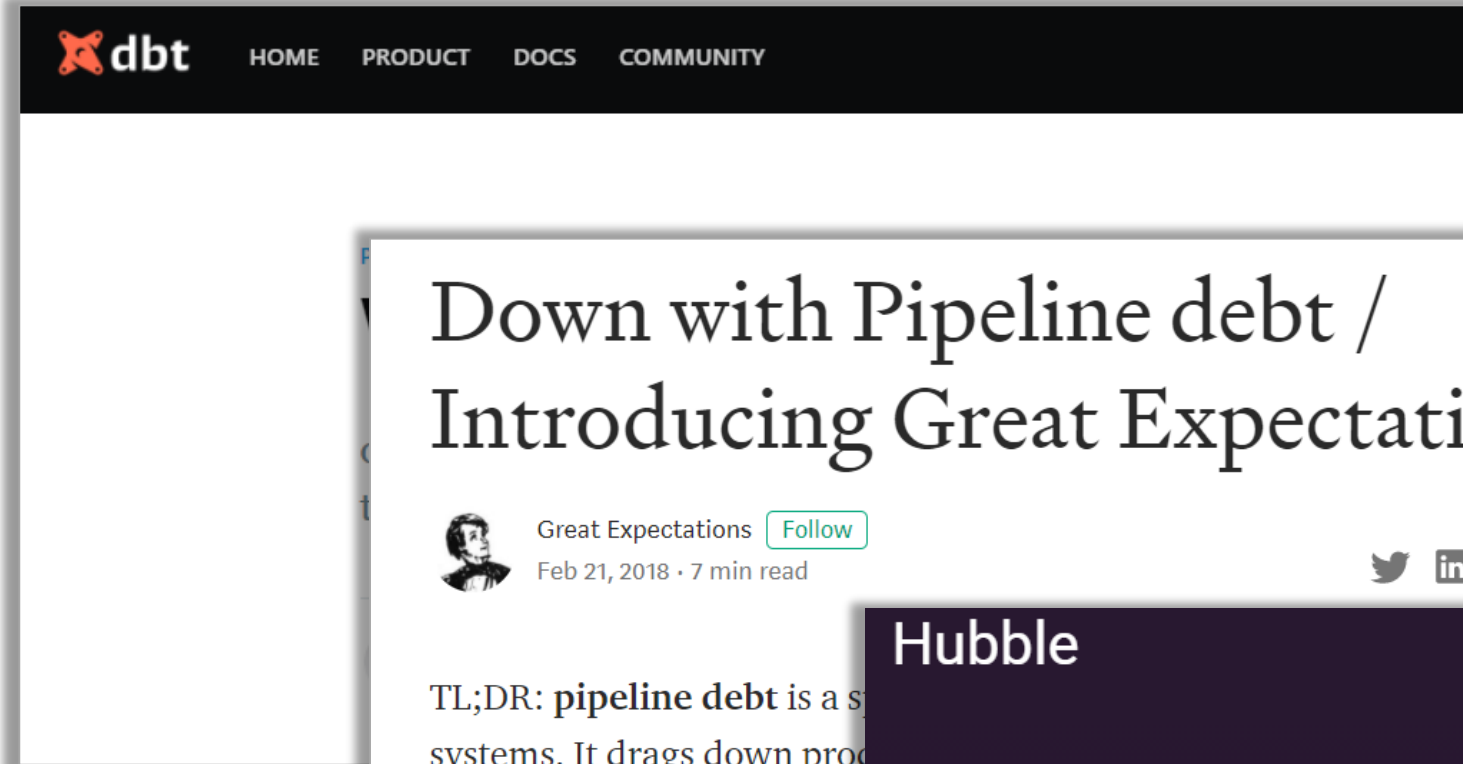
Impala/Hive - Spark - Presto - Druid

Kafka - Pulsar - Flink - Nifi - Beam - Spark

TXT - CSV - JSON - Avro - Parquet

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

Transzformáció és adatminőség



The screenshot shows the top navigation bar of the dbt website with links for HOME, PRODUCT, DOCS, and COMMUNITY. Below the navigation is the main content area of a blog post. The title is 'Down with Pipeline debt / Introducing Great Expectations'. The author is 'Great Expectations' with a 'Follow' button. The date is 'Feb 21, 2018' and the estimated reading time is '7 min read'. There are social media sharing icons for Twitter, LinkedIn, Facebook, and a bookmark icon. The main text of the article is partially visible, starting with 'TL;DR: pipeline debt is a s... systems. It drags down pro... best way to beat pipeline de... tests, which are applied to... of compile or deploy time)... source tool that make it eas... and join our slack!'.

Hubble

Automated monitoring of all your data metrics

Hubble monitors your data warehouse for data quality issues and alerts you when things go wrong.

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 Blob Storage - Google Cloud Storage

Apache Airflow

Airflow is a platform created by the community to
program



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

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 Blob Storage - Google Cloud Storage

Metaadat-kezelés



MARQUEZ

Collect,

vi

Amundsen — Lyft's data discovery & metadata engine



Mark Grover [Follow](#)
Apr 2, 2019 · 10 min

This post introduces about technical arch

In order to increase scientists at Lyft, we metadata engine. Co-[Roald Amundsen](#)), providing a search i

DataHub: A generalized metadata search & discovery



Mars Lan August 14, 201

Co-authors

Editor's no

February 2

As the open LinkedIn's

How We Improved Data Discovery for Data Scientists at Spotify



February 27, 2020
Published by Andrew Maher



Lexikon

Search



[Contribute research](#)

[FAQ](#)



Welcome to Lexikon.

Find your insights.

Overview

Marquez is an open source ecosystem's metadata. It m global visibility into job run management, and much m

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 Blob 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 Blob Storage - Google Cloud Storage

Thank You

