Evolution of a modern cloud-based data lake

Viacheslav Inozemtsev
viacheslav.inozemtsev@zalando.de

O’Reilly Software Architecture Conference
Berlin, 07.11.2019
OUTLINE

Introduction of Zalando

Why data lake?

What is a data lake?

Evolution of the data lake

Future
Who am I?

Viacheslav Inozemtsev

- 2010 - Specialist in Applied Mathematics and Computer Science at the Tomsk State University, Russia
- 2014 - Master of Computer Science at the University of Bonn, Germany
- Total of 8 years of experience as Data and Software Engineer
- Last 3 years at Zalando
Introduction of Zalando
Introduction of Zalando

- 29M active customers
Introduction of Zalando

- 29M active customers
- 14000 employees
Introduction of Zalando

- 29M active customers
- 14000 employees
- 2000 engineers and data scientists
Introduction of Zalando

- 29M active customers
- 14000 employees
- 2000 engineers and data scientists
- 200 engineering teams
Introduction of Zalando

● 29M active customers
● 14000 employees
● 2000 engineers and data scientists
● 200 engineering teams
● Variety of major data systems
  ○ Messaging Bus
  ○ BI Data Warehouse
  ○ Google Analytics platform
  ○ Custom datasets
Why data lake?
Why data lake?
Why data lake?
Why data lake?
Why data lake?
Why data lake?
Why data lake?
Why data lake?
Why data lake?
Why data lake?
Why data lake?

- To enable *sharing* of the *data* so, that it is:
Why data lake?

- To enable **sharing** of the **data** so, that it is:
  - easy to publish and consume
Why data lake?

- To enable **sharing** of the **data** so, that it is:
  - easy to publish and consume
  - compliant
Why data lake?

● To enable **sharing** of the **data** so, that it is:
  ○ easy to publish and consume
  ○ compliant
  ○ secure
Why data lake?

- To enable **sharing** of the **data** so, that it is:
  - easy to publish and consume
  - compliant
  - secure
  - cost-efficient
What is a data lake?
What is a data lake?

- Central **exchange** system
  - connects producers and consumers of the data
  - defends data from malicious misuse or accidental leaking
What is a data lake?

- **Central exchange system**
  - connects producers and consumers of the data
  - defends data from malicious misuse or accidental leaking

- **Central big data system**
  - provides as much different data as possible
  - has to be fast and easy to use
What is a data lake?

● Central **exchange** system
  ○ connects producers and consumers of the data
  ○ defends data from malicious misuse or accidental leaking
  ○ *open vs hide!*

● Central **big data** system
  ○ provides as much different data as possible
  ○ has to be fast and easy to use
What is a data lake?

- Central **exchange** system
  - connects producers and consumers of the data
  - defends data from malicious misuse or accidental leaking
  - open vs hide!

- Central **big data** system
  - provides as much different data as possible
  - has to be fast and easy to use
  - scale vs performance!
Evolution of the data lake
Evolution of the data lake - Stage 1: Inception

AWS S3
Evolution of the data lake - Stage 1: Inception

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hive Metastore</td>
<td></td>
</tr>
<tr>
<td>AWS S3</td>
<td></td>
</tr>
</tbody>
</table>
Evolution of the data lake - Stage 1: Inception

- Data Catalog
- Hive Metastore
- AWS S3
Evolution of the data lake - Stage 1: Inception

- Event Bus
- BI Data Warehouse
- Google Analytics
- Custom datasets

Data Catalog
Hive Metastore
AWS S3
Evolution of the data lake - Stage 1: Inception

- Event Bus
- BI Data Warehouse
- Google Analytics
- Custom datasets

Data Catalog
Hive Metastore
AWS S3

Direct S3 access

presto

Evolution of the data lake - Stage 1: Inception

- Event Bus
- BI Data Warehouse
- Google Analytics
- Custom datasets

Data Catalog
Hive Metastore
AWS S3

presto
Direct S3 access
Evolution of the data lake - Stage 1: Inception

- Event Bus
- BI Data Warehouse
- Google Analytics
- Custom datasets

Data Catalog
Hive Metastore
IAM assume role
AWS S3

Direct S3 access
presto
Evolution of the data lake - Stage 1: Inception
Evolution of the data lake - Stage 2: Optimization
Evolution of the data lake - Stage 2: Optimization
Evolution of the data lake - Stage 2: Optimization
Evolution of the data lake - Stage 2: Optimization
Evolution of the data lake - Stage 2: Optimization
Evolution of the data lake - Stage 2: Optimization
Evolution of the data lake - Stage 2: Optimization
Evolution of the data lake - Stage 2: Optimization

Delta Lake is an open-source storage layer that brings ACID transactions to Apache Spark™ and big data workloads.
Evolution of the data lake - Stage 2: Optimization

- Event Bus
- BI Data Warehouse
- Google Analytics
- Custom datasets

<table>
<thead>
<tr>
<th>Data Catalog</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hive Metastore</td>
</tr>
<tr>
<td>IAM assume role</td>
</tr>
<tr>
<td>AWS S3</td>
</tr>
</tbody>
</table>

- databricks
- presto

Direct S3 access
Evolution of the data lake - Stage 2: Optimization
Evolution of the data lake - Stage 3: Revolution

- Event Bus
- BI Data Warehouse
- Google Analytics
- Federated custom datasets

Data Catalog
Hive Metastore
IAM assume role
AWS S3

Direct S3 access

https://martinfowler.com/articles/data-monolith-to-mesh.html
Evolution of the data lake - centralized
Evolution of the data lake - federated
What is data lake?

- Central **exchange** system
  - connects producers and consumers of the data
  - defends data from malicious misuse or accidental leaking
  - **open vs hide!**

- Central **big data** system
  - provides as much different data as possible
  - has to be fast and easy to use
  - **scale vs performance!**
What is data lake?

- Federated **exchange** system
  - connects producers and consumers of the data
  - defends data from malicious misuse or accidental leaking
  - open vs hide!

- Federated **big data** system
  - provides as much different data as possible
  - has to be fast and easy to use
  - scale vs performance!
What is data lake?

- Federated **exchange** system
  - connects producers and consumers of the data
  - defends data from malicious misuse or accidental leaking
  - **open vs hide!**
  - resolved by **centralized governance**
- Federated **big data** system
  - provides as much different data as possible
  - has to be fast and easy to use
  - **scale vs performance!**
What is data lake?

- Federated **exchange** system
  - connects producers and consumers of the data
  - defends data from malicious misuse or accidental leaking
  - **open vs hide!**
  - resolved by **centralized governance**

- Federated **big data** system
  - provides as much different data as possible
  - has to be fast and easy to use
  - **scale vs performance!**
  - resolved by **decentralized ownership**
Future
Future

- Make data lake fully federated
Future

- Make data lake fully federated
- Abstract physical storage into logical layer of datasets
Evolution of a modern cloud-based data lake

Viacheslav Inozemtsev
O’Reilly Software Architecture Conference
Berlin, 07.11.2019

viacheslav.inozemtsev@zalando.de
www.zalando.com
jobs.zalando.com/tech