From Legacy to Cloud

An End to End Data Integration Journey

Max Schultze
max.schultze@zalando.de

@mcs1408

02-05-2019
Max Schultze

- Data Engineer
- MSc in Computer Science
- Took part in early development of Apache Flink
- Retired semi-professional Magic: the Gathering player
# TABLE OF CONTENTS

- Legacy Analytics
- Zalando’s Data Lake
- Data Integration Pipelines
- Processing Platform
- Lessons Learned
Legacy Analytics
Legacy Analytics

DWH
Legacy Evolving

Messaging Bus

Data Lake
Zalando’s Data Lake
Zalando’s Data Lake

Ingestion

- Event Bus
- Data Center
- DWH
- Web Tracking

Storage

Serving
Zalando’s Data Lake

Ingestion
- Event Bus
- Data Center
- DWH

Storage
- Metastore

Serving
- Web Tracking

DWH: Data Warehouse
Storage: Data Storage
Serving: Data Serving
Zalando’s Data Lake

Ingestion

- Event Bus
- Data Center
- DWH
- Web Tracking

Storage

- Metastore

Serving

- Ad-Hoc Querying
- Processing Platform
- Data Catalog
Data Integration Pipelines
Data Integration Pipelines

Ingestion

Event Bus

Data Center

DWH

Web Tracking

Serving

Ad-Hoc Querying

presto

Processing Platform

databricks

Data Catalog

Storage

Metastore
Event Data Pipeline

NAKADI

File Explosion

New File Queue

Data Lake Raw

Nakadi Archiver

Dead Letter Queue
Data Integration Pipelines

Ingestion
- Event Bus
- Data Center
- DWH
- Web Tracking

Storage
- Metastore

Serving
- Ad-Hoc Querying
  - presto
- Processing Platform
  - databricks
- Data Catalog
Relational Data Pipeline

Invoke Lambda every N minutes

Poll events

Ready for extraction
- Check running datasets
- Trigger new SFs

Extraction SF

Event Bus
Delta Table Format

- Table format on top of Parquet
- Additional metadata written to cloud storage
Delta Table Format

- Table format on top of Parquet
- Additional metadata written to cloud storage
- Files are only “active” when committed to delta metadata
Delta Table Format

- Table format on top of Parquet
- Additional metadata written to cloud storage
- Files are only “active” when committed to delta metadata

⚠️ No expensive file listing
Delta Table Format

- Table format on top of Parquet
- Additional metadata written to cloud storage
- Files are only “active” when committed to delta metadata

⚠️ No expensive file listing

✔️ Transactionality of readers
Delta Table Format

- Table format on top of Parquet
- Additional metadata written to cloud storage
- Files are only “active” when committed to delta metadata

⚠️ No expensive file listing
✅ Time traveling
✅ Transactionality of readers
Delta Table Format

- Table format on top of Parquet
- Additional metadata written to cloud storage
- Files are only “active” when committed to delta metadata

⚠️ No expensive file listing
✔️ Transactionality of readers
✔️ Time traveling
✔️ Data deletion for GDPR
Delta Table Format

- Table format on top of Parquet
- Additional metadata written to cloud storage
- Files are only “active” when committed to delta metadata

- No expensive file listing
- Time traveling
- Transactionality of readers
- Data deletion for GDPR

Delta Lake is now open source: https://delta.io
Processing Platform
Processing Platform - Technical Setup

Databricks AWS

Databricks Control Center

Zalando AWS

Zalando Databricks

Team Z

Team Y

Team X

Zalando Data Center

DWH

Data Lake
Processing Platform - Technical Setup

- Databricks AWS
  - Databricks Control Center
- Zalando AWS
  - Zalando Databricks
  - Team Z
  - Team Y
  - Team X
- Zalando Data Center
  - DWH
- Data Lake
Processing Platform - Technical Setup

- Databricks AWS
  - Databricks Control Center

- Zalando AWS
  - Zalando Databricks
  - Team Z
  - Team Y
  - Team X

- Zalando Data Center
- DWH

- Data Lake
# Processing Platform - Organizational Setup

## Introduction to Databricks

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Solution Architect
- Office Hours
Processing Platform - Organizational Setup

Introduction to Databricks
- Solution Architect
- Office Hours

Initial Setup
- Inner Source Configuration
Processing Platform - Organizational Setup

- Introduction to Databricks
  - Solution Architect
  - Office Hours

- Development Phase
  - Office Hours
  - Guest Developer

- Initial Setup
  - Inner Source Configuration
Processing Platform - Organizational Setup

- **Introduction to Databricks**
  - Solution Architect
  - Office Hours

- **Initial Setup**
  - Inner Source Configuration

- **Development Phase**
  - Office Hours
  - Guest Developer

- **Productionizing**
  - 24/7 Support
Federated Data Lake

Metadata Layer

Processing Platform
- datalake
- presto

Data Lake Storage

AWS RDS

Amazon Redshift
Lessons Learned
Tales to Tell

Go Serverless
Tales to Tell

Go Serverless

Automation, Automation, Automation...

- CI/CD
- Infrastructure as Code
- Inner Source Configuration
Tales to Tell

Go Serverless

Automation, Automation, Automation...

- CI/CD
- Infrastructure as Code
- Inner Source Configuration

The Biggest Data Integration Challenges are not Technical
From Legacy to Cloud

An End to End

Data Integration Journey

Max Schultze
max.schultze@zalando.de

@mcs1408