Building data-driven applications with SAP Data Hub and Amazon Web Services

Dr. Lars Dannecker, Steffen Geissinger
September 18th, 2018
Enterprise data landscapes are growing increasingly complex

LANDSCAPE CHALLENGES

GOVERNANCE
Lack of security and visibility. Who changed the data? What was changed? Who is accessing it?

LIMITED TOOLS
Lack of enterprise readiness. High effort to productize complex data scenarios across data landscape

MISSING LINK
Between Big Data and Enterprise Data. Data is kept in silos across the enterprise.
Example: Advanced Fitness Tracker Analysis

Heart rate, battery speed, distance

Sales figures
Revenue, Margin
Customer Groups
Location, Details

Big Data

Enterprise Data (e.g. ERP)

Transactions
Customer Data
Sales Order

Intelligent Enterprise

SAP Data Hub
Machine Learning
Data Mining
Advanced Applications
Intelligent Assistants

© 2018 SAP SE or an SAP affiliate company. All rights reserved. | PUBLIC
<table>
<thead>
<tr>
<th>Category</th>
<th>Number of Customers in Million</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casual Jogger</td>
<td>10.2</td>
</tr>
<tr>
<td>Persistent Runner</td>
<td>5.1</td>
</tr>
<tr>
<td>Enthusiastic Sprinter</td>
<td>3.7</td>
</tr>
<tr>
<td>Professional Athlete</td>
<td>0.6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Total Revenue with... in Million</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casual Jogger</td>
<td>$745</td>
</tr>
<tr>
<td>Persistent Runner</td>
<td>$1,739</td>
</tr>
<tr>
<td>Enthusiastic Sprinter</td>
<td>$914</td>
</tr>
<tr>
<td>Professional Athlete</td>
<td>$812</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>Average Revenue per Customer in $</th>
</tr>
</thead>
<tbody>
<tr>
<td>Casual Jogger</td>
<td>73</td>
</tr>
<tr>
<td>Persistent Runner</td>
<td>341</td>
</tr>
<tr>
<td>Enthusiastic Sprinter</td>
<td>247</td>
</tr>
<tr>
<td>Professional Athlete</td>
<td>1,354</td>
</tr>
</tbody>
</table>

3 Most Popular Shoes:
- Casual Jogger: Aerosol III, Speedy, Relax (3 Series)
- Persistent Runner: Speedy, Aerosol III, Glide-X 4700
- Enthusiastic Sprinter: Aerosol III, Glide-X 4700, Speedy
- Professional Athlete: RLAG 100-200, Glide-X 4700, Speedy

Not Assigned: 7.8 million customers
SAP Data Hub – Integrate, orchestrate, process

Enterprise Systems:
- SAP ERP
- SAP S/4HANA
- SAP BW/4HANA

SAP Data Hub:
- Data Discovery and Metadata Governance
- Orchestration & Data Pipeline
- Connectivity, Integration, Ingestion

Distributed Data Systems:
- Data Hub Runtime
- Hadoop
- Cloud Storage
- Machine Learning

Holistic Landscape Orchestration
Flow-based data processing
Enterprise and 3rd party connectivity
## Data Hub is the answer for enterprises’ modern data management challenges

<table>
<thead>
<tr>
<th>Topics</th>
<th>Current challenges</th>
<th>Data Hub’s value proposition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Visibility</td>
<td>Growing difficulty in managing and orchestrating large volume of data coming from SAP and non-SAP system</td>
<td>Seamlessly orchestrates data from SAP and non-SAP system and provides <strong>360 degree view of ALL company data</strong></td>
</tr>
<tr>
<td>Data Quality</td>
<td>Constant need to improve data quality by cleansing and resolving inconsistencies</td>
<td>Offers one cross-landscape <strong>control center to monitor and improve data quality</strong></td>
</tr>
<tr>
<td>Data Innovation</td>
<td>Complexity in utilizing data from SAP and non-SAP systems for Machine Learning training and IoT use cases</td>
<td>Efficiently streams and processes data from all sources to <strong>enable new Machine Learning and IoT use cases</strong></td>
</tr>
<tr>
<td>Data Cost</td>
<td>Increasing storage and compute costs due to growing data volumes (organically or via M&amp;A)</td>
<td><strong>Optimizes data costs</strong> by eliminating data duplications and data movement</td>
</tr>
<tr>
<td>Data Compliance</td>
<td>Enforcing multiple corporate and regulatory data policies is becoming a burden and risk for enterprise IT</td>
<td>Manages all <strong>data compliance &amp; governance</strong> policies of a company in one central location</td>
</tr>
</tbody>
</table>
SAP Data Hub – Architectural overview

Connected Systems
- SAP S/4HANA
- SAP BW/4HANA
- SAP Data Services
- SAP LT Replication Server
- SAP HANA
- SAP Cloud Applications (API-driven)
- Native Integration with AWS Offerings
- Open connectivity for 3rd party & open source

Distributed Runtime
- Relational
- Time-Series
- Graph
- Document
- SAP VORA Engines

Pipelines & Workflows
- Scripting (JS, Python)
- Built-in Connectors
- Custom Operators
- Flow-based Applications
- Metadata Catalog
- Connectivity

Metadata & Applications
- Scheduling
- Profiling & Discovery
- Application Services

SAP Data Hub System Management (based on SAP HANA)
- Multi-Tenancy
- User & Access Management
- Content Lifecycle Management
- Cluster Management
- Diagnostics

Data Storages
- Cloud / On-Premise
- Object Stores
  - Amazon S3
- Hadoop
  - HDFS (optional)
- SAP Data Hub Adapter
- VORA Spark Extensions

© 2018 SAP SE or an SAP affiliate company. All rights reserved. | PUBLIC
**SAP Data Hub data pipelines concept**

**SAP Data Hub Pipelines**

- **Pipelines = Computation Graphs**
- Execution environments for operators:
  - **Docker Containers** as execution environments
  - Operators Groups → running in same Kubernetes Pod
  - Groups with multiplicity → parallel execution
- **Sample operators**: Amazon Kinesis Consumer, JavaScript Operator, S3 File Producer, HTTP Client
Easily design complex pipelines

- **Navigation Pane**: Use this pane to access operators, graphs, repository, and the types.

- **Pipeline (Graph) Editor**: Use this editor to create a pipeline (graph) with one or more operators.

- **Editor Toolbar**: The graph editor includes a toolbar, which you can use to perform operations on the graph, for example, to save and execute a graph.

- **Status Pane**: You can use this pane to monitor the status of the graph execution, trace messages and view various logs.
Building Data Driven Applications
Why should I even think about hybrid?

ERP Data  Customer Data  Sales Data

Customer Data Center

High Value Enterprise Data

Machine Learning

Cloud Storage

IoT/Stream Processing

Advanced Analytics
AWS & Data Hub – Jointly processing for improving your analysis

SAP Data Hub

- **Connection Management**
  - Landscape Orchestration
  - Connect and organize your entire system landscape to get a holistic overview.

- **Metadata Catalog**
  - Metadata Governance
  - Get insights into the characteristics, profiles and data models of data residing in connected systems.

- **Data Preparation**
  - Data Refinement
  - Define the necessary steps to refine and shape the data with respect to your specific processing steps.

- **Data Joins**
  - Joint Processing
  - Join data from different sources and continue processing them together.

- **Data Pipelines and Transformations**
  - Advanced flow-based applications
  - Conduct advanced transformations or apply machine learning by building powerful pipelines.

**Data Hub Storage**

- Disk-Based Data Hub Storage
  - EBS Volumes
- Data Hub Data Lake
  - S3 Storage

**Source**

- 3rd Party Cloud
  - Cloud Storage
  - Messaging Services
  - Cloud Databases
  - Machine Learning
  - Internet of Things

**SAP Systems**

- SAP HANA
- SAP BW/4HANA
- ERP, S4/HANA
- SAP Applications
- SAP Cloud Platform
- Data Services

**Target**

- 3rd Party Cloud
  - Cloud Storage
  - Messaging Services
  - Cloud Databases
  - Machine Learning
  - Internet of Things

- SAP Systems
  - SAP HANA
  - SAP BW/4HANA
  - ERP, S4/HANA
  - SAP Applications
  - SAP Cloud Platform
  - Data Services

© 2018 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC
AWS & SAP Data Hub Major use case scenarios

Rapidly integrate and leverage new data sources
Big Data warehouse use case
- Acquire data from enterprise and cloud sources
- Combine structured and unstructured data
- Seamlessly move selected data sets across landscapes

Understand real-world performance
Internet of Things use case
- Combine streaming sources with static enterprise sources
- Support for high-velocity data ingestion and processing
- Scale-out ingestion processing and pipelining

Machine Learning and Predictive Analytics
Data Science use case
- Apply machine learning to any data set
- Operationalize and automate Machine Learning processes
- Wide variety support
**Modern landscapes: data perspective**

- **Enterprise data warehouse**
  - Analytical modeling
  - Joins/union
de-normalization
  - Matching/
duplicate check
  - Metadata extraction
  and generation
  - Enrichment
  - Cleansing
  - (Re-)
formatting
  - Anonymization/masking
  - Parsing
  - Filtering
  - Search
  - Data
  validation
  - Data stream

- **Big Data**
  - Refining Insights out of (Big) Data
    - Structuring the unstructured
    - Handling large volumes of data
    - ETL or DWH are not the answer
    - Data formats and granularity
    - Data streams and flexible structures
    - Apply logic to the data, not data to the logic
    - Integrated analytics
    - Early insights on all levels
    - Automation needed

- **Raw data**
  - #123
  - 10101
  - $%&$?
Vision: Intelligent Data Warehouse

- Close Integration between enterprise systems and cloud services
- High automation and minimal Modeling
- Data Lake as primary high volume and computation persistency
- Scalable Storage and Data Processing capabilities in Cloud / On-Premise
- Data Processing beyond OLAP with ML / Predictive Analytics etc.
Orchestration with SAP Data Hub
Data Workflows – Data replication to the cloud

- Orchestration via **pipelines** chaining multiple operations and execution engines
- Operators are managed and running as part of SAP Data Hub
- The heavy lifting / logic execution happens either:
  - **Internally** – In the operator container
  - **Externally** – In the connected system
Realizing the Big Data Warehouse for customers

SAP DATA HUB – Runtime

Raw Data (S3 Data Lake)

Data Pipelines & Flows

SAP Vora

SA/P Data Services

Data Flows

Amazon EKS

SAP DATA HUB

Meta Data Catalog

Workflow Orchestration

SAP Data Services

Integration

Analytics Model

Structured Data

Amazon Redshift (via Data Services)

S/4HANA

Refined Data

Big Data Processing

Ingestion

ERP

Databases

Social Media

Kinesis

Amazon MQ
Demo
IoT – Understanding Real-World Performance
SAP Data Hub - Use case scenario
Internet of Things (IoT)

Data Hub and AWS: Core capabilities
- Unite streaming data (sensors) with enterprise data (business metadata)
- Event-based execution and processing
- Scaling to 1,000s of pipelines in parallel at any time
- Automate, design, and run all data processes

Examples
- Information from Internet-enabled devices
- Customer demographics
- Supply chain information
- Granular product usage information
- Apply the concept of ‘Digital Twins’ to data streams, enabling customers to test outcomes and impacts of potential actions
Vision: End-to-End IoT Architecture with SAP Data Hub and AWS IoT

Data Sources Collect Publish / Subscribe Process (Stream / Batch) Store Analyze / Serve

IoT Edge

AWS IoT Edge

Thing

AWS IoT Core

File Gateway

IoT Platform

SAP Data Hub (Pipeline Engine)

Ingestion Pipeline / Orchestration / Streaming

Spark Custom Code Rules

Op1 Op 2 Op 3 Op 4

Amazon Kinesis

Events

Raw Data

SAP Vora

Processing Engines

Relational Time Series Graph Document

Disk Persistence (Warm Data)

Mass Storage

SDA

Raw Data Store

HDFS (e.g. EMR)

SAP HANA

In-Memory Engine (Hot Data / Aggregates)

Applications

IoT Application

Analytic Application

Business Solution (e.g. ERP, CRM)

IoT Enterprise

© 2018 SAP SE or an SAP affiliate company. All rights reserved. ǀ PUBLIC
Example: Fitness Tracker Analytics
Machine Learning – Intelligent Data Processing
SAP Data Hub - Use case scenario
Machine Learning and Predictive Analysis

Customer Challenges
- Leveraging machine learning algorithms beyond the data science team
- Difficulty in monetizing and scaling out machine learning across an enterprise

What we need to provide
- Apply machine learning and predictive algorithms to any data set
- Operationalize ML processes
- Insert machine learning and predictive processing to any scenarios

Examples
- Insurance industry risk profiling
- Credit analysis and automated scoring models
- Machine failure prediction leading to automated preventative maintenance
Example: Advanced analytics for the chemical industry in Germany

Churn Analysis

Train Model Pipelines
- Read + prepare data from connected system
- Fit and deploy ML-models in Repository

Serve Model Pipelines
- Consume ML-models from Repository
- Expose prediction services via REST endpoint and WebUI

Analytical Tools
Stitching it Together to Build Data-Driven Applications
Designing data-driven applications with Data Hub and AWS

Data-driven applications

IoT  Machine Learning  Analytics / BW  ...

SAP Data Hub

Metadata Management

Orchestration

Processing & Pipelines

Integration & Ingestion

Distributed Datalake (In-Memory, Disk, S3)

HANA / Vora integration

ABAP Integration

Workflow

Cloud Integration API

Connectors (open & native protocols)

- Amazon MQ
- S3 Storages
- EMR / HDFS
- REST APIs
- AWS Databases
- Amazon Analytics
- Machine Learning
- SCI for process integration
- SAP Event Bus

Amazon and distributed data sources

SAP Applications

© 2018 SAP SE or an SAP affiliate company. All rights reserved. | PUBLIC
Why Amazon is the perfect partner for data-driven applications

<table>
<thead>
<tr>
<th>Managed Kubernetes Service</th>
<th>Proven Data Lake Capabilities</th>
<th>Advanced Analytics and Machine Learning</th>
<th>Established Customer Base</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple deployment</td>
<td>Virtually unlimited storage</td>
<td>Huge variety of machine learning</td>
<td>Larges cloud provider</td>
</tr>
<tr>
<td>Operations and SLAs</td>
<td>High availability &amp; durability</td>
<td>Well-established analytics</td>
<td>Best established for</td>
</tr>
<tr>
<td>Easy auto scaling</td>
<td>Security &amp; Compliance</td>
<td>Native Tensorflow integration</td>
<td>enterprise customers</td>
</tr>
<tr>
<td>Certified K8s conformance</td>
<td></td>
<td></td>
<td>Suited for high security</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>projects (GovCloud)</td>
</tr>
</tbody>
</table>
Vision: Hyper-flexible scaling with Amazon Fargate
Thank you.

Contact information:

Dr. Lars Dannecker  
Big Data Architect P&I Big Data  
lars.dannecker@sap.com

Steffen Geissinger  
Big Data Architect P&I Big Data  
steffen.geissinger@sap.com