SAP Data Hub

Innovative and powerful data integration in today’s complex world

PUBLIC

Christoph Streubert, SAP
April 3rd, 2019
Agenda

• Introduction & Challenges SAP Data Hub addresses

• Product Overview & Values

• Architecture

• Use Case Examples

• Custom Content

• Next Steps
We are facing Enterprise Data landscapes that are diverse and increasingly complex

LANDSCAPE CHALLENGES

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

**PIPELINES?**
Too hard to refine and enrich heterogeneous data across multiple systems.

**DATA SHARING?**
Integration is manual, point-to-point, painful, and slow.
The Challenge
Bridging Two Different Worlds

Innovations

Focus:
Key features & Proof Points

Fast
Explorative
Latest technology

Agile
Easy to pilot

Business Applications

Focus:
Completeness & Operations

Sustainable
Compliant
Extendable
Maintainable
User Friendly
Integrated
Easy to roll-out

Latest technology

Stable
The Challenge
Bridging Two Different Worlds

Innovations
- Machine Learning
- Image Recognition
- Recommendation
- Hadoop
- Tensorflow
- Kafka
- Spark
- S3
- Web Services
- AWS

Business Applications
- Dashboards for BI & Analytics
- BW
- CRM
- ERP
- MII
- Geo Spatial
- Web Shop
- Logistics
There are Challenges to uniting the landscape
We need to overcome silos, cost and complexity to drive better operations and insight

**Missing link**
Expand and Integrate Big Data and enterprise data

**Lack of enterprise-ready security and governance**
Connect securely Across complex landscapes

**Many limited tools = high effort, cost, time**
Create new digital applications with simplified data management
Delivering intelligent data for the intelligent enterprise
Right data to the right users with the right context at the right time

Disparate Data Sources → Data Landscape Management
Harness the power of exponentially growing data

Data Pipelines
Accelerate intelligent data delivery

Data Governance
Compose pervasive data consumption strategies

IT Administrator → Data Admin → Data Engineer → Data Scientist → Data Steward → Business Analyst

Data Consumption
SAP Data Hub: Data orchestration solution
End-to-end, all-in-one solution

End-to-end
Data Lakes
Applications
Databases
Data Warehouses
Data Marts
Cloud Datastores
Third-Party Data Services

All-in-one
Data Ingestion
Data Discovery
Data Refinement
Data Enrichment
Data Sharing
Data Governance

Disparate Data Sources → Data Landscape Management → Data Pipelines → Data Governance → Data Consumption

Data Landscape Management
Harness the power of exponentially growing data

Data Pipelines
Accelerate intelligent data delivery

Data Governance
Compose pervasive data consumption strategies
SAP Data Hub: All-in-one solution for enterprise data orchestration

Data Landscape Management
Ingestion & connectivity
Security & policy management
Monitoring & deployment

Data Pipelines
Modeling data pipelines & workflows
Data enrichment
Data preparation & quality
Distributed data pipeline processing

Data Governance
Data discovery
Data profiling
Metadata cataloging
Data landscape management
Most trusted, open, and flexible end-to-end data landscape management

View the full data landscape, orchestrate all data movement, and understand your data and how it impacts data models and results.

Integration and connectivity
Leverage existing connections and integration tools, while adding new connections easily and flexibly

Security and policies
Manage security settings, data accessibility, and data policies to ensure appropriate data governance and security

Monitoring
Comprehensive view of systems and assets across the full landscape
Data pipelining
All-in-one reusable, automated data pipelines

Create powerful data pipelines that cleanse, conform, transform, ingest, refine, orchestrate, prepare, and enrich information from a variety of sources across the organization – while leaving the data where it resides.

Define the data operations process
Reuse existing code and libraries through data pipelines consisting of several predefined and customizable operations.

Data preparation & quality
Transform, cleanse, match, and consolidate data, and understand the impact of quality problems on all downstream systems and applications.

Distributed data pipeline processing
Process data at the source and manage data operations across the data landscape.
SAP Data Hub Features

Modeler (Data Pipelines): Graphically build powerful data pipelines using a variety of reusable components for connectivity (HDFS, Kafka, SAP HANA, REST, etc.) and processing (Spark, JavaScript, Python, etc.)
The Pipeline Modeler provides a large variety of predefined operators together with their environments for productive use cases:

- Connectors to **messaging systems** (Kafka, MQTT, NATS, WAMP, Google Pub/Sub)
- Connectors to **store** and **read data** (HDFS, S3, File, NFS, GCS, WASB, etc.)
- Operators for **RESTful** clients and services (e.g. APIs, App Services)
- Connectors to **databases** and **enterprise apps** (SAP Business Suites, S/4 HANA, BW/4 HANA, SAP HANA, SAP Vora, etc.)
- Operators for **data processing** (Javascript, Python, Go, etc.)
- Operators for **process execution** (stateful and stateless)
- Operators for **Spark & R**
- Operators for **machine learning** (Tensorflow, MLF)
- Operators for **image processing** (OpenCV)
- Operators for digital **signal processing**
- Operators for **type conversion**
Data Integration & Processing
Operators for Managed Data Sources

New connectivity operators:
• *ABAP ODP Consumer (PILOT Version)
  • See the SAP Pilot Note 2731192 for more technical details
• DB2 SQL Consumer
• DB2 Table Consumer
• SQL Server SQL Consumer
• SQL Server Table Consumer
• MySQL SQL Consumer
• MySQL Table Consumer
• Google BigQuery Table Consumer
• Google BigQuery SQL Consumer
Data governance
Delivering the right data to the right users with the right context at the right time

Gain a clean view of your data landscape and its interconnections, no matter where the data lives, by centralizing the orchestration not the data.

Data profiling
Analyze data in order to clarify the structure, relationships, and deviation rules of data, and understand anomalies

Data discovery
Profile, view, and expose the data of all connected systems, and evolve data models more quickly in a highly visual environment

Metadata cataloging
Define, govern, and manage your metadata assets across enterprise systems with disparate sources
Unify data for your solution to achieve **scalable visibility and control**

- **Single system view** – for data pipelining, orchestration, monitoring, and governance

- **No centralization of data** – no mass data movement to a single data store

- **Distributed native processing** – executes pipeline activities quickly, where the data resides
Architecture overview of SAP Data Hub

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

Cloud Stores
- AWS S3, GCP GCS, Azure ADL & WASB

Hadoop
- HDFS (optional)

SAP Data Hub Adapter

SAP Vora Spark Extensions

Connected Systems
- SAP S/4HANA
- SAP BW/4HANA
- SAP Data Services
- SAP LT Replication Server
- SAP HANA
- Databases
- SAP Cloud Applications (API-driven)

Open connectivity for third-party & open source
The on-premise challenge

- What is Kubernetes?
- Operate Kubernetes cluster
- Operate elastic block storage
- Operate load balancers
- Operate docker infrastructure
- Operate object store

- Operate private cloud infrastructure
SUSE CaaS Platform

Key Features

- **Workload scheduling** places containers according to their needs while improving resource utilization.

- **Non-disruptive Rollout/Rollback** of new applications and updates enables frequent change without downtime.

- **Service discovery and load balancing** provides an IP address for your service, and distributes load behind the scenes.

- **Health monitoring** and management supports application self-healing and ensures application availability.

- **Application scaling** up and down, accommodates changing load.

https://www.suse.com/products/caas-platform/
IoT data ingestion and orchestration with SAP Data Hub
Integrate and process disparate data from messaging systems and high-volume cloud storages

- Tackle the challenge of integrating and analyzing vast quantities of raw data and events from disparate semi-structured sources with low-level semantics and no business context

- Solve the point-to-point challenge of distributed heterogeneous environments spanning messaging systems, cloud storages, SAP data management solutions, and enterprise apps

- Event-driven pipelines scaling to executions of many pipelines in parallel, at any time
Data science & machine learning data management with SAP Data Hub

Prepare data from all source to increase effectiveness of AI/Machine learning algorithms

- Integrate variety of data sources with open data landscape management
- Process ML models leveraging many SAP and non-SAP engines within the same tool
- Quickly and safely operationalize ML outcomes back into enterprise processes
- Dramatically decrease the time needed to collect, refine, and orchestrate data
Intelligent data warehouse with SAP Data Hub
Rapidly integrate and leverage new data sources

- Acquire new data sources with previously siloed data from traditional data warehouses, data marts, enterprise applications, and Big Data stores
- Combine all types of sources including structured and unstructured data, and enable a large variety of processing on them
- Seamlessly process large data sets across highly distributed landscapes and close to the data source, moving only high-value data
Governance and data cataloging with SAP Data Hub
Understand and secure your data

- Crawl through data stores to gather valuable metadata and store it in a centralized information catalog
- Profile source data to gain a deeper understanding of the data to create meaningful data pipelines
- Move to centralized data access and control for all orchestration, data refinement, scheduling, and monitoring
Big Data Warehouse
Customer Architecture Example

**Access**
- SAP Analytics Cloud
- SAP Analytics Cloud
- SAP PA / Spark
- 3rd Party

**Store & Process**
- Master Data
- Hadoop (HDFS)
- SAP HANA
- SAP VORA
- S3

**Orchestration & Data Refining**
- MODELLING
- TRANSFORM
- EXTRACT
- FEDERATE
- LOAD
- JOIN
- FILTER
- CLEANSE
- LOCK-UP
- SCRIPT
- MASK
- ANYMONIZE
- PARSE
- STREAM
- COPY
- BATCH
- SAP HANA or BW/4HANA
- SAP Data Hub

**Example Scenario**
- Combine refined big data with enterprise data and corporate master data
- Extract or federate data into SAP HANA or BW/4HANA
- Ingest Data into S3 as Landing Zone for data
- Orchestrate and schedule all related processes
- Implement transformations and data pipelines
- Harmonize data structures and look up of reference data
- Execute operations on large data volumes
- Automation of complex data science processes and decision making based on data in-flight
Internet of Things
Customer Example

How to understand customer behavior and to drive insightful decisions for customers and R&D.

- Customer Example
  - Smart appliances sending sensor data which is processed to get insight into customer usage patterns which in turn drives R&D investments
  - ~6 million devices, 16 TB data volume per day

Solution with SAP Data Hub
  - Refine business value from data ingestion to enterprise applications
  - Visual modeling environment
  - Governance and Data Management
  - Orchestration, refining and scheduling to define automated data driven processes enabling real-time action taking and decision making
SAP Data Hub Custom Content
SAP Data Hub ISV positioning options and framework

- SAP Developed
  - Available on ISV pricelist
  - Provided to ISV/OEM partner as-is
  - Custom development dependent on use case and business plan

SAP Data Hub as-is

- Data Hub
  - SAP delivered functionality

SAP Data Hub Custom content

- Data Hub execution Framework
  - (Docker, Kubernetes, Pipeline Modeler)

- Partner developed, priced and supported content
- Technically validated by SAP
- Compatible to new and existing Data Hub installations
- Available to customers through SAP Content Hub
- Sold through SAP App Center

Custom development dependent on use case and business plan.
SAP Data Hub – extended functionality

Integration categories

• **Custom Operator**
  – Extend pre-defined base operators using powerful runtimes, e.g. Python, Go, Javascript, and R
  – Develop new operators with own runtimes, e.g. Process Executor

• **3rd Party Docker Images**
  – Provide execution environments for operators
  – Incorporate 3rd party libraries and executables

• **Pipeline Templates**
  – Parameterizable scenario templates
  – (Sub)graphs called in other graphs
  – Industry solutions as complete pipelines with data provisioning requirements

• **Embedded Applications (planned 2H 2019)**
  – Think additional tiles on Data Hub launchpad

• **SAP Data Hub as-is**
  – SAP delivered functionality

• **SAP Data Hub Custom content**
  – Partner developed, priced and supported content
  – Technically validated by SAP
  – Compatible to new and existing Data Hub installations
  – Available to customers through SAP Content Hub
  – Sold through SAP App Center

• **Data Hub execution Framework**
  – (Docker, Kubernetes, Pipeline Modeler)
SAP Data Hub Extensibility

Extension Points

- Tenant Applications managed by System Management
- Pipeline Modeler
  - Custom Operators
  - Built-in Connectors
  - Workflow Operators
- Metadata Explorer
  - Metadata Catalog
  - Profiling / Discovery
- Flow Agent
- Connection Management
- Embedded System Application
- SAP Vora Engines
  - Relational
  - Time-Series
  - Graph Document
- Database Tools
- Spark on Kubernetes

SAP Data Hub System Management
- Multi-Tenancy Management
- User & Access Management
- Content Lifecycle Management
- Cluster Management
- Repository
- SAP HANA (internal)

Connected Systems
- SAP Applications
- SAP Cloud Applications (API-driven)
- Databases
- Public Clouds
- 3rd Party / Open Source

*SAP HANA (internal) has been added as a standalone box in the diagram*

3rd party Docker Images

Docker Registry

for SAP Data Hub Pipeline Modeler

© 2019 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

*This is the current state of planning and may be changed by SAP at any time without notice.*
SAP Data Hub Extensibility
Building custom Operators and Docker Images

- The Pipeline Modeler provides a couple of predefined **Base Operators**

- **Custom Operators** derived from the Base Operators and can be extended with custom
  - Scripts
  - Parameters
  - Input- and Output Ports
  - Auxiliary Files
  - Documentation

- **Dockerfiles** describe the container runtime for operators and are chosen based on **Tags**

- The Pipeline Modeler stores all related artifacts in the Pipeline Modeler **Repository**

*This is the current state of planning and may be changed by SAP at any time without notice.*
SAP Data Hub Content Lifecycle
Example – Create Custom Operator and Pipeline using the Operator
SAP Data Hub Content Lifecycle
Example - Export Project as Solution
SAP Data Hub Content Lifecycle
Example - Export Project as Solution

```json
{
  "name": "Hello World",
  "version": "v1"
}
```
SAP Data Hub Extensibility
Open for Partner and Customer Content Delivery

Data Hub’s Extensibility Options
Partners and customers can add own content based on different layers:

- **Pipelines Templates:**
  - Parameterizable scenario templates
  - (Sub)graphs called in other graphs

- **Operators:**
  - Extend pre-defined base operators using powerful runtimes, e.g., Python, Go, Javascript, and R
  - Develop new operators with own runtimes, e.g., Process Executor

- **Dockerfiles:**
  - Provide execution environments for operators
  - Incorporate 3rd party libraries and executables

Data Hub Content Delivery

- Custom Data Hub Content can be packed into a **SAP Data Hub Solution Archive**
- SAP Data Hub Solutions can be defined, exported and imported via
  - SAP Data Hub System Management UI
  - vCTL Command Line Tool*
- **SAP Data Hub Solution Structure:**

```
vsolution
  ├── vrep
  │   ├── vflow
  │   │   ├── dockerfiles
  │   │   └── graphs
  │   └── operators
  │       └── path
to package
  └── vsolution.json
```

SAP Data Hub Content Marketplace:

- Planned Marketplace and Delivery channel for SAP Data Hub Partner Content via **SAP API Business Hub** in combination with **SAP App Center** for monetization*

A win-win partnership:

- Partners can reach more customers with their services, while SAP Datahub Hub expands its reach with relevant content
- Our partners understand the customer use cases and can better support with domain-specific content by generalizing the specific issue with a broad solution

*This is the current state of planning and may be changed by SAP at any time without notice.*
Get started with SAP Data Hub

Try SAP Data Hub, developer edition

Subscribe to SAP Data Hub videos YouTube channel

Take free SAP Data Hub Course

Technical Documentation, help.sap.com

Find out more about SAP Data Hub sap.com/datahub

partneredge.sap.com/data-hub

Try SAP Data Hub, Trial edition

Develop with SAP Data Hub, and add developer license to your services (DEV_7019390)
Appendix
Unified user experience
One central entry point to all services and applications
SAP Data Hub metadata explorer

Redesigned and improved user experience

Build up catalog to get insight into your companies metadata

Centralized location

Browse connections
- Monitoring
- Metadata catalog
- Search datasets
- Publications
- Labels
Data integration and connectivity
Unified connectivity framework

The connectivity framework (Flowagent) served as the underlying infrastructure with the goal to rapidly grow and enhance the native connectivity and integration functionalities:

**Metadata Services** (Browsing, Profiling, Data Preview)
- Hadoop (HDFS)
- Cloud Object Storages (AWS S3, GCP GCS, Azure Data Lake, WASB)
- Oracle*, ABAP/ODP*, OData*

**Connection Operators** (Consumer, Producer)
- HDFS, S3, GCS, ADL, WASB
- Oracle**, ABAP/ODP**, OData**
- Support custom adapters

**Spark Code Generation**
- HDFS

*profiling is planned in future release  **producer is planned in future release
Holistic landscape visibility, control, and metadata management

Understand the overall data landscape and ensure data security

• Describe and organize the landscape within a single environment
• Enforce access policies to secure data dynamically from source to destination
• Mask and anonymize data to protect sensitive sources

Rapidly manage and automate your data models

• Use a single entry point for monitoring, scheduling, and landscape health
• Build a repository of reusable and extensible metadata models
Access and harmonize information from a variety of sources

Work across a diversity of systems and data

- Open architecture to work across a broad array of data sources: Cloud or on-premise, Big Data, or enterprise data (SAP or non-SAP (third-party))
- Common orchestration of data and processes to overcome system boundaries
- Accelerated development time with visual modeling

Magnify the impact of your data

- Ensure data quality with predefined operations to cleanse, process, and transform data
- Process large volumes of data where they occur to gain the insights you need
Simple and scalable approach to manage and monitor data flows

Ease the cost and effort of data management and integration

• Connect quickly to your diverse, distributed data landscape
• Easily include more data sources and targets
• Empower more users to manage data and data pipelines with a single, user-friendly solution

Respond quickly to opportunities to create and improve data flows

• Identify opportunities to connect systems and information to gain new insights
• Accelerate and automate data processes and remove unnecessary process redundancies
• Improve the effectiveness of data results by resolving data quality issues or friction points
Consume all data and enable data-driven apps and processes

Create powerful data pipelines
• Use distributed Big Data processing
• Take advantage of serverless computing paradigms for radical scale
• Incorporate powerful third-party libraries, such as TensorFlow

Build data streams to enable data-driven processes
• Establish data streams to react to data changes instantly
• Embed complex algorithms into existing data flow by using any code or script
Unified metadata catalog to gain visibility about landscape wide data assets

- Easily govern and manage metadata assets across enterprise system disparate sources
- Discover, understand, and consume information about data with the ability to synchronize, share, and perform lineage and impact analysis
- Answer related information requests without browsing through multiple systems and repositories, or touching various data models
- Support non-domain experts in evaluating data quality and the impact of changes