Data Quality Challenges in Enterprise Big Data Landscapes

Eric Simon
Chief Scientist, SAP P&I Big Data

Data Quality Theory and Applications Workshop, Lyon, December, 2017

Run Simple

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

Outline

Introduction

Data Quality Assessment

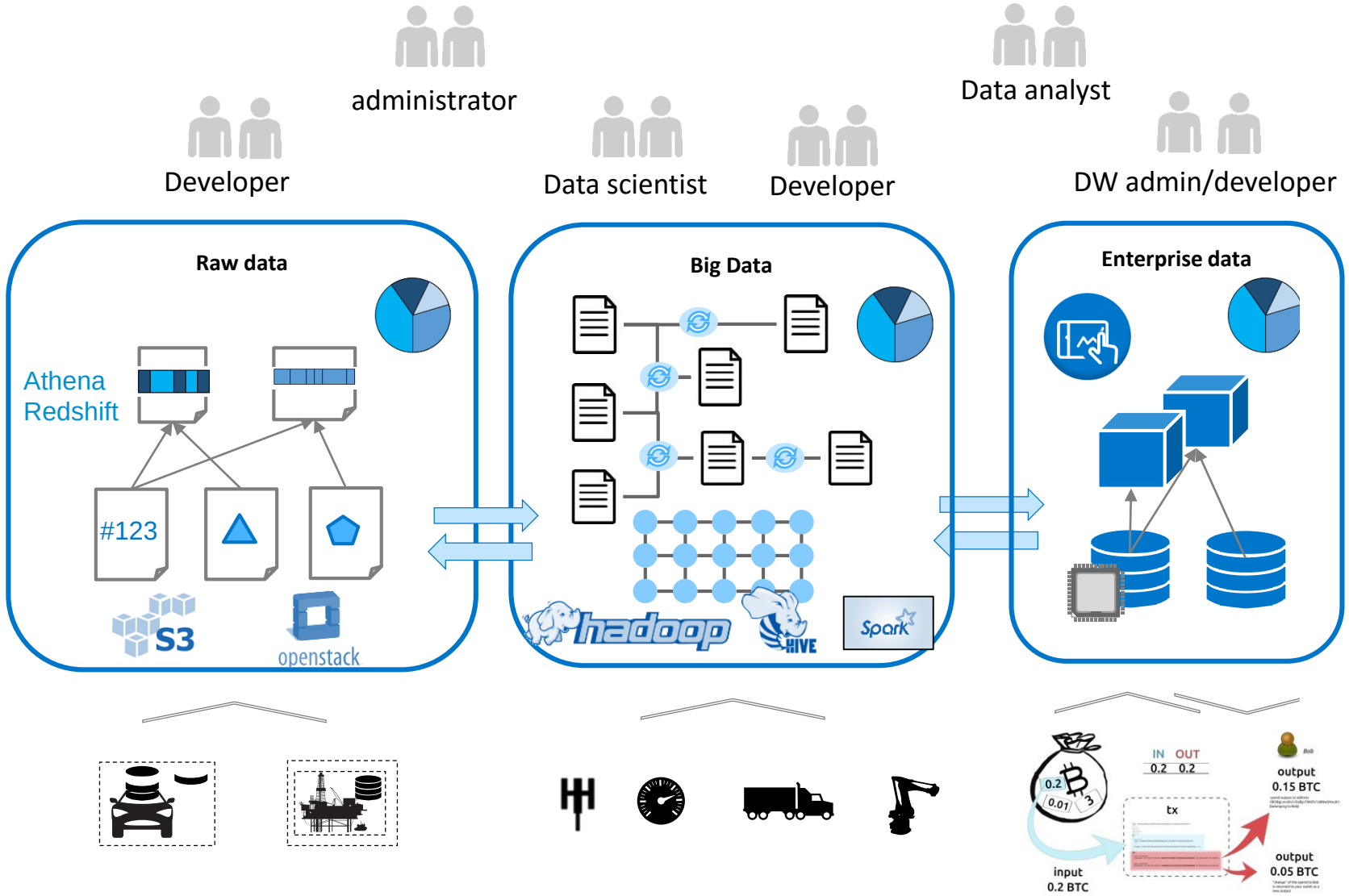
Continuous Data Quality Monitoring

Data Repair

Conclusions

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

Evolution Towards Enterprise Big Data Landscape



Diversity of

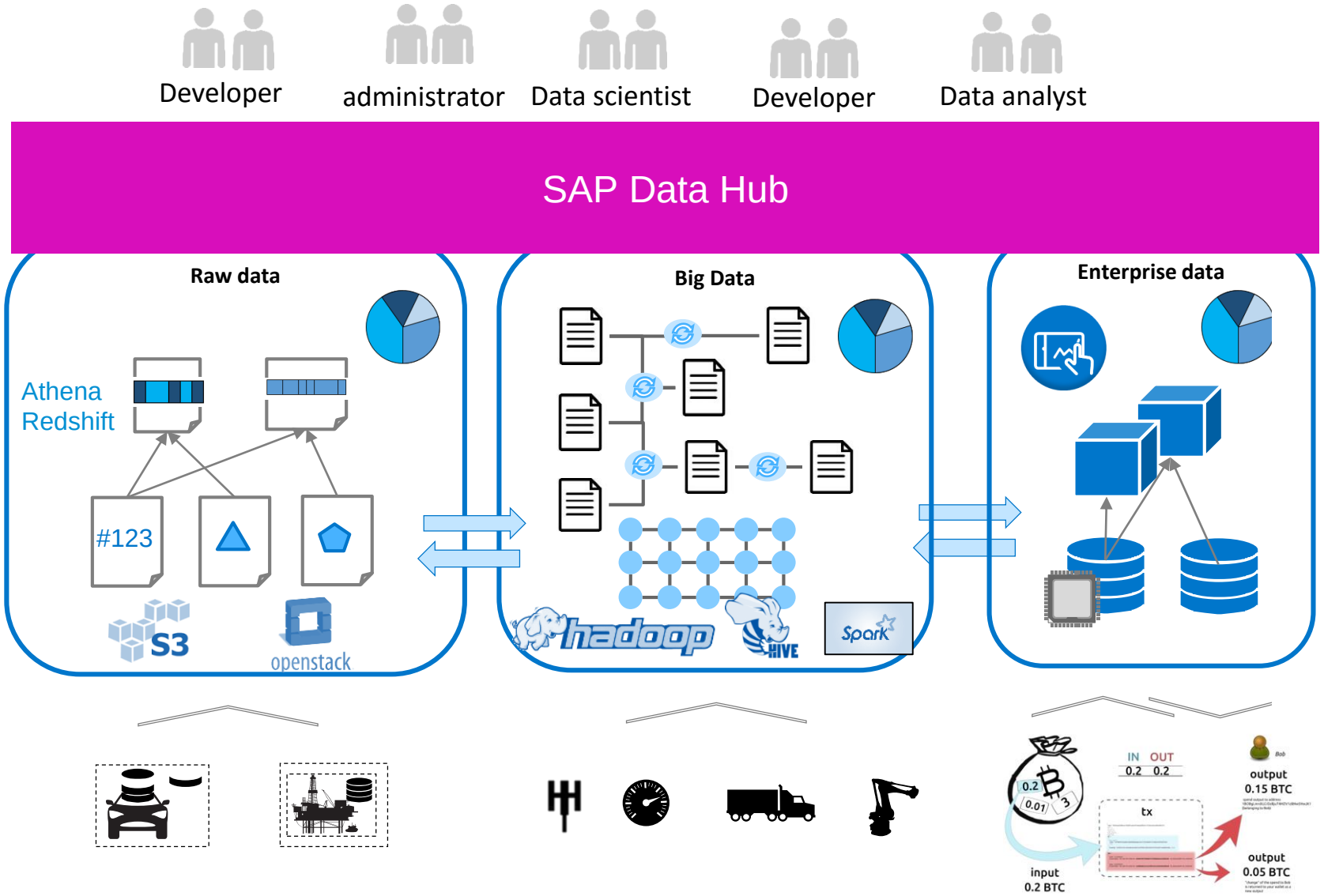
- Technology stacks
- Technical requirements
- Consumer persona

Data processing spread

- Within and between stacks
- Functional overlap
- Data volume and velocity

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

Evolution Towards Enterprise Big Data Landscape



Diversity of

- Technology stacks
- Technical requirements
- Consumer persona

Data processing spread

- Within and between stacks
- Functional overlap
- Data volume and velocity

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

SAP Data Hub

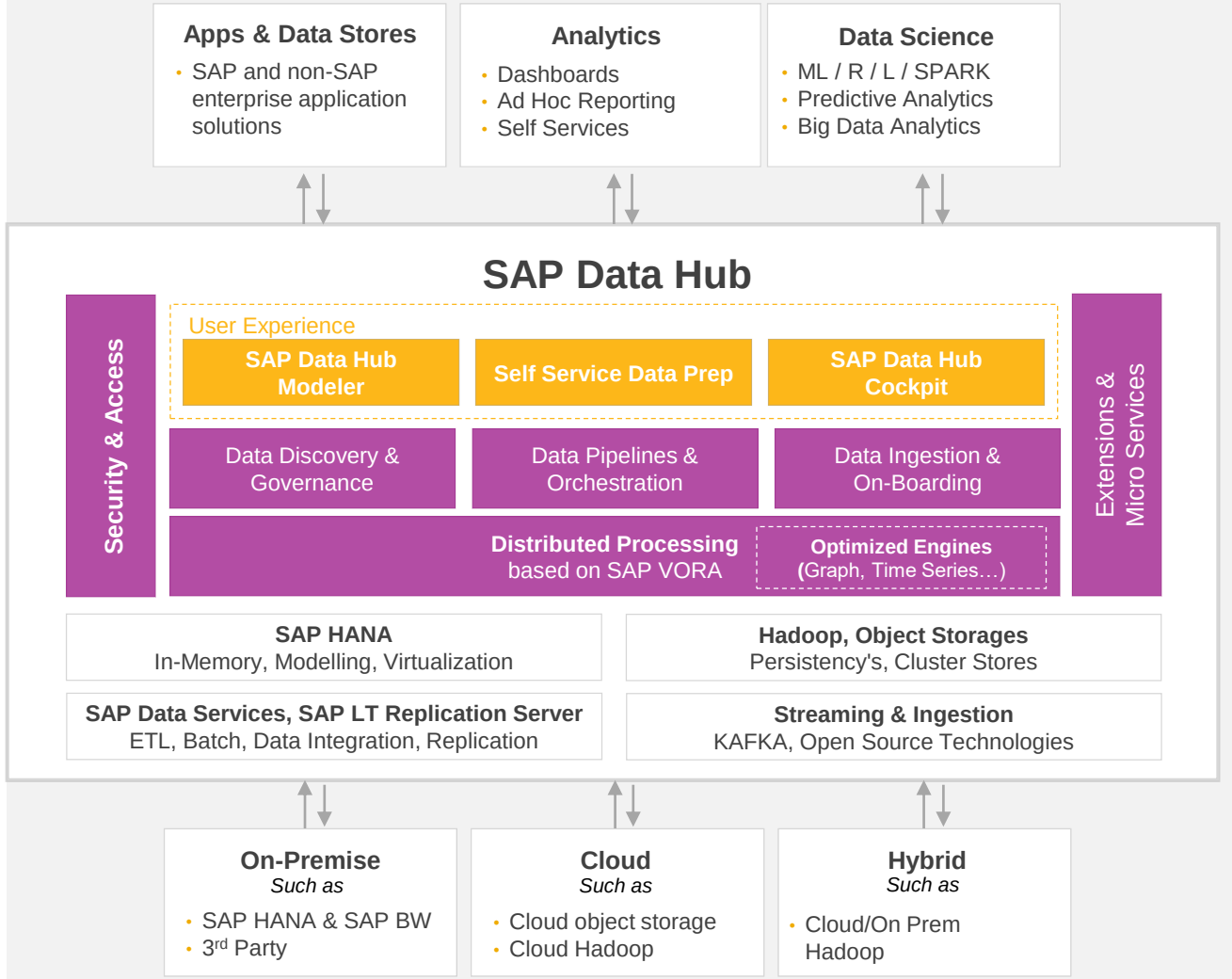
Overview and Key Functionality

Efficient Data Sharing – Leverage existing connections and integration tools, while adding new connections easily and flexibly. Access on-premise or cloud data sources.

Centralized Governance – metadata repository of information stored in the connected landscape. Offering profiling, quality monitoring, data lineage, and search.

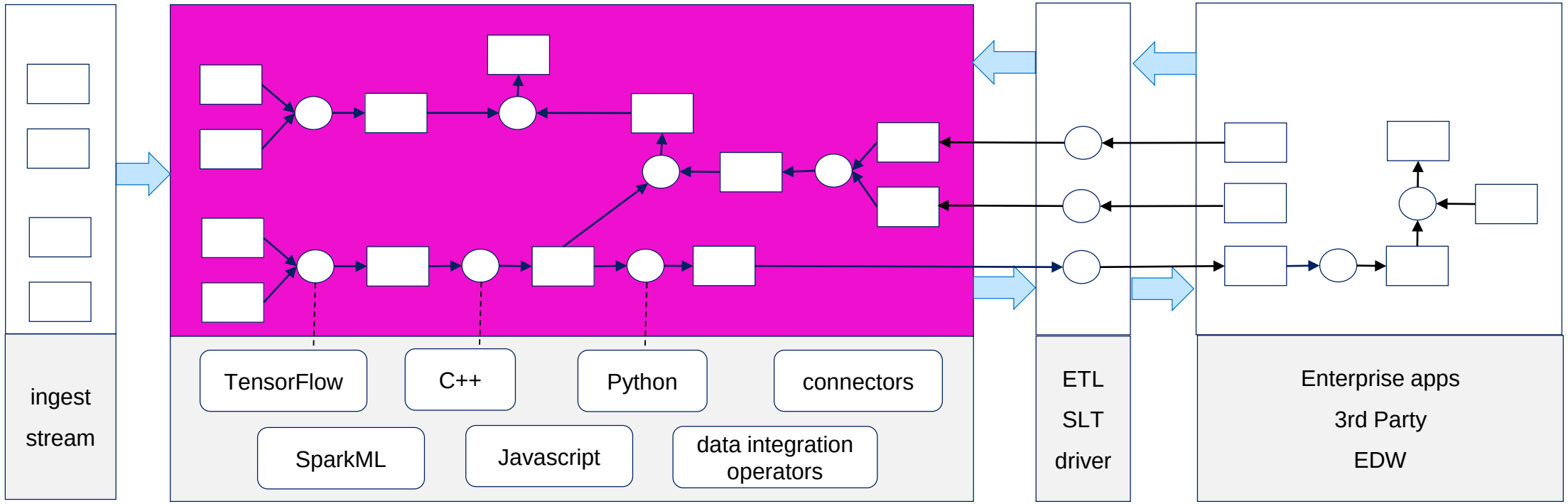
Data Pipelines – flow based applications consisting of reusable and configurable operations, e.g. data transformations, native code execution, trained models, connectors. Executes where data resides.

Workflows – orchestrate processes across the data landscape, e.g. executing data pipelines, triggering SAP BW Process Chains, SAP Data Services Jobs and many more



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

SAP Data Hub
Positioning of Data Pipelines



- Dataflow programming model with reusable and configurable operators; extensibility provided by integration with external frameworks and libraries (e.g., Tensorflow, image processing SDK)
- Data pipeline is one slice of the end-to-end data processing

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

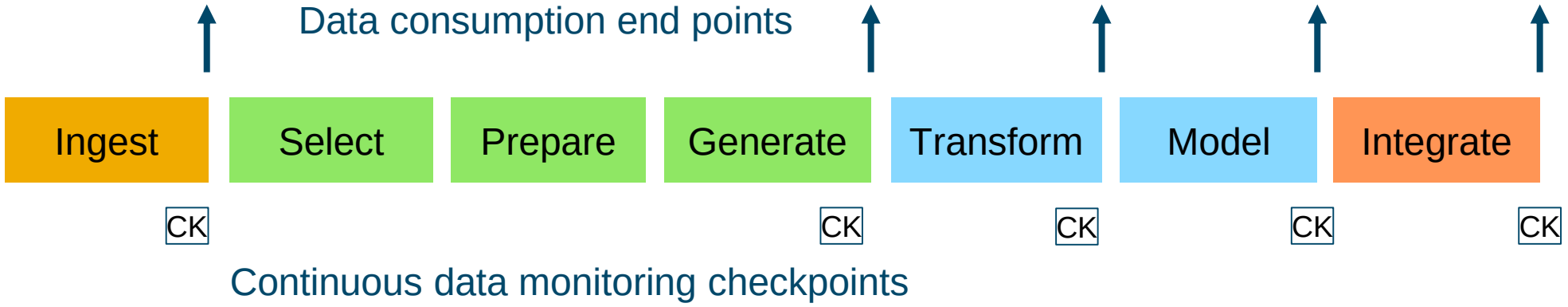
Data Quality in Big Data Landscapes

IT cannot protect anymore the “perimeter” !

- Data go beyond EDW and DM, good and bad data co-exist, GIGO principle is not sustainable
- Data quality must get closer to the data consumers ⇒ “make data quality everyone’s business”
- Consumers = data scientists, data analysts, power business users

Consequences

- Provide means to assess data content according to user’s goal and fix observed data
- Continuously monitor the quality of data flowing through data pipelines
- Assist the user in repairing the data as needed and where needed



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

Enable Users to Assess and Fix Data Content

Assess native datasets

- Note 1: user has domain expertise and knows his/her objective
- Note 2: existing “agile data preparation” tools already provide powerful means to profile individual columns (e.g., analyze "nulls", patterns, ranges, outliers, data type detection) and suggest data transformations

Data type detection

- Data type detection methods rely on reference data that are limited to party data. Collecting and maintaining other data has no sufficient ROI for now. How to collect and maintain good reference domain data at affordable price?
  - Typical case is “codification” fields that are ubiquitous (e.g., thousands of code fields in ERP database, technical codes used in signal data)
  - User understands business data but is less exposed to the meaning of codes and what they identify
  - Codes can be universally or locally unique (e.g., product id, room id, SSN, …)

Detect entities

- Most datasets contain associations between entities. How to detect entities and their meaning? Impacts:
  - Distinguish dimensions from metrics (usable by outlier detection methods and analytics)
  - Consolidate information about entities occurring in diverse datasets and detect inconsistencies ⇒ ad hoc master data
  - Help discovering related datasets (useful for feature engineering and enterprise analytics)

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

Enable Users to Assess and Fix Data Content

Assess native datasets

Multi-column data dependencies

- Draw user attention on “interesting” data dependencies (e.g., Approximate FD ⇒ Prune “incidental” AFDs)
  - Detection of entities can reduce complexity, user-provided critical variables (e.g., days of therapy, quantity dispensed)
- Ease understanding of data dependencies by means of examples in an interactive manner
  - Find examples that are valid for 1 dependency but irrelevant for as many others as possible
  - Select counter-examples that most probably indicate an error ⇒ suggest data transformations

Effective data sampling to improve iterative process

- Get a first data sample: assess, harmonize, clean, and enrich the data ⇒ generate data pipeline
  - First sample should be as “representative” as possible (e.g., as with stratified sampling). Suggest/ask initial categorical variables (known to expert depending on objective) – AFD can also be used initially
- Get a new data sample, run data pipeline and assess quality of results
  - subsequent samples should satisfy certain data dependencies found during previous iterations
  - counter-examples of data dependencies (e.g., AFD) can be included
  - When nulls are produced when LOJ-ing with other datasets, more distinct values can be included to seek non-nulls

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

Enable Users to Assess and Fix Data Content

Assess native datasets

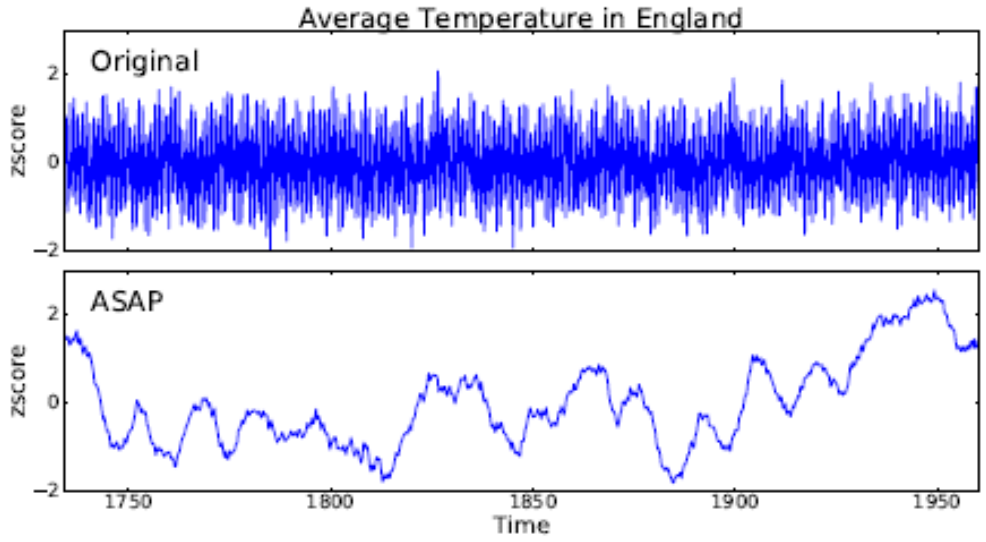
Temporal data

- Profiling of time-dependent data (time series)
  - Correct behavior may not be known
  - Detect significant changes in behavior across stream windows
  - Detect missing values
- Visualization of anomalies
  - e.g., automatic smoothing techniques in ASAP
- Stability of the dataset over a long time period
  - History of past behaviors (e.g., instrument measures)

Origin and Usage of data

- Who created the data? Where is the data used ?
- For which analytics? Created by who?
- Are there transformed versions of the data?

  ⇒ all of this requires data lineage and tagging capabilities



from ASAP – VLDB 2017

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

Enable Users to Assess and Fix Data Content

Assess transformed datasets

Assess data transformations that produced the data

- Data transformations may “drop” data (e.g., filter old data) or destroy the characteristics of data (e.g., interpolation of signal data, elimination of outliers). Is it more pertinent to reuse data pipeline’s input data ?

All of this requires understanding where the data come from, how it was produced. E.g., does the pipeline acts as filter with respect to its input, where do the features come from? ⇒ Data lineage (both schema and instance level)

Assess datasets input/output of machine-learning pipelines

- Does a dataset result from an ML model M? Reusing it silently will cause a “hidden debt” for M
- Is a dataset a stable version of a model output? ⇒ assure stable dependencies
- Does dataset contain observed or learned features? requires tracking the origin of features
- Does the dataset contain unmaintained legacy features? ⇒ avoid underutilized data dependencies
- Will reusing a dataset result in increasing the “pipeline jungle” ? (e.g., add more complexity to a Data Prep pipeline)

All of this requires understanding data dependencies in complement to dataset and feature annotations

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

Enable Users to Assess and Fix Data Content

Assess transformed datasets

Reuse analytics to complement a dataset

- Add analytics using a complement dataset (complement via LOJ on common attributes)
- Control is needed to avoid low quality results
  - Semantic relationships between datasets must be trusted (e.g., do they share same dimensions)
  - Pre-aggregation of an analytic dataset may not fulfill aggregable properties of measures (e.g., stock measures)
  - Pre-aggregation can be ambiguous (e.g., aggregate population by city and state in Table 2)
  - Heterogeneous units and currencies must be handled
- Relates to “summarizability” problem

Country	State	City	Population
USA	*Texas*	*Dublin*	3 654
USA	*Texas*	*Paris*	25 171
USA	*Maine*	*Paris*	5 183
France	–	*Paris*	2 200 000
Panama	–	*Paris*	1 070

Table 2: Population_Detail

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

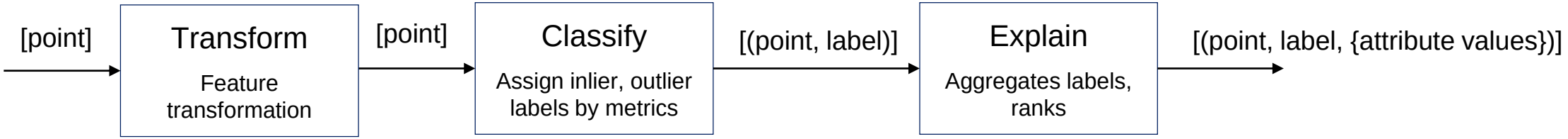
Continuously Monitoring Data Quality

Traditional Data Quality project deployment

- Data quality process is created to curate source data (e.g., master data)
- Data quality monitoring is put in place to control data quality continuously at all levels
  - Cleanse and transform data as they flow through data pipelines
  - Output rejected data for inspection: based on rule checking (order of 100 – 1,000 rules) or cleansing operator's logic

Discover unanticipated data anomalies when high velocity precludes manual inspection

- Generic classification and interpretation operators for fast data stream analysis
- Default classification operator: density-based classification using robust statistical estimation (MAD, MCD estimators)
- Adaptable Damped reservoir to retrain estimators



from Macrobase – SIGMOD 2017

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

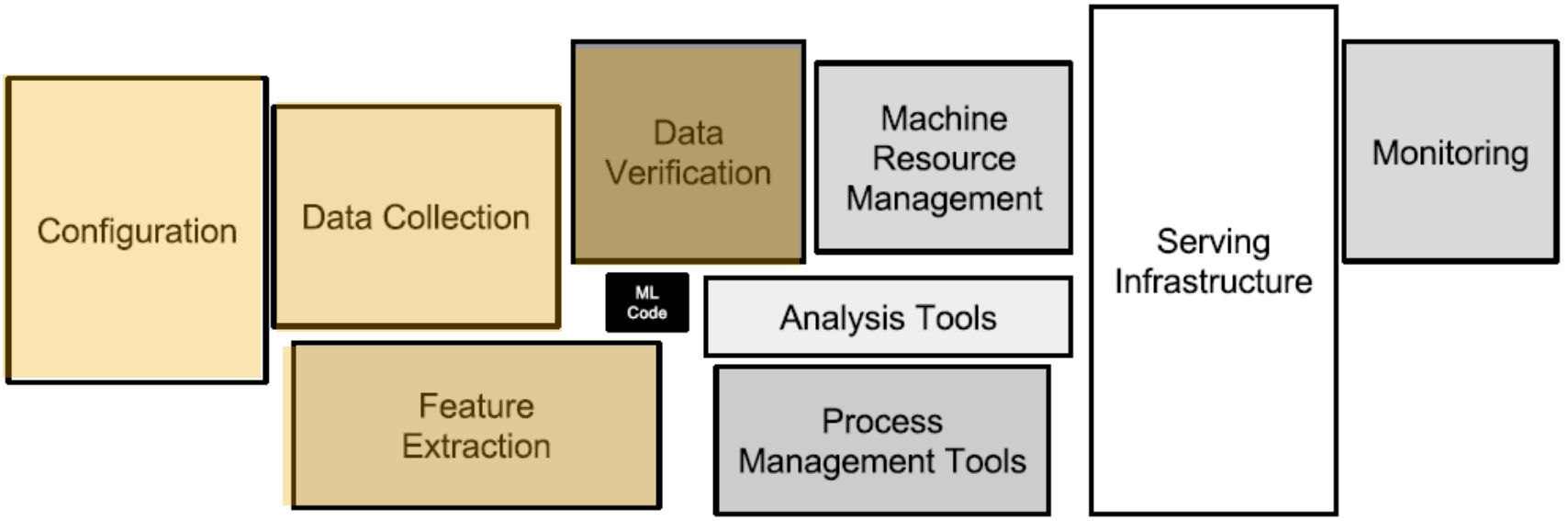
Continuously Monitoring Data Quality

Silent failures depending on data dependencies

- Typically occurs when predictive models are used in data pipelines, requires analysis of data dependencies and statistical monitoring of data (datasets, attributes)
- Feature obtained from old (outdated) data due to lack of refresh
  - Behavior will decay gradually
- Change of coverage for a feature or change of value distribution
  - E.g., drop from 90% to 60% of non-nulls in a feature, may cause misbehavior
  - Feature entanglement in a model
- Change in the domain of a predicated variable
  - E.g., sentiment analysis changes from 2 to 5 distinct values, can break analytics precision
  - Prediction bias: check that the distribution of predicted labels is the same as the distribution of observed labels

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

Continuously Monitoring Data Quality



From “Hidden technical debt in ML systems” – Google report 2014

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

Repairing the Data

General repair data quality guidelines

- No automatic correction/repair ! Don’t loose the data! ⇒ human-driven remediation
- Notify errors, group them and present to users for review
  - for predefined rules (e.g., bad address), repair actions can be suggested; in others cases we can’t
- Support root cause analysis ⇒ errors can be detected lately in the data pipeline
- Fix the errors : where and how to fix ?
- Analyse impact of repair action ⇒ impact on downstream pipeline

Data provenance for root cause analysis

- Existing (eager) techniques for data pipelines have high space and time overhead for data provenance capture
  ⇒ Need for mixed approaches that exploit schema-level lineage and “disclosing” aproaches
- Useful techniques: replay and exclusive-replay
  – Must be carefully handled and depends on “correctness” of data provenance
- Other techniques like mining data lineage to discover changing patterns (e.g., map fan-out in secure networks)

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

Repairing the Data

What and where to fix?

- Update the pipeline
  - Change data transformations to remedy the data (add/remove filters, add column transformations)
  - Branch the pipeline at some step and create an alternate version, either to address special cases (contingency pipeline) or to eliminate unwanted dependency on a shared pipeline
- Create intermediate and versioned data
  - Versioned data bring stability (limit variations and unexpected changes)
  - Cloning provides insulation from downstream transformations in other pipelines using the same dataset (e.g., create a curated version – could be virtual)
- Add rules for earlier error detection

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

Conclusions

Enterprise Big Data landscape is changing the game

- Data processing is spread across multiple stacks with a large end-to-end scope
- Non traditional data manipulations beyond relational-like operations, including ML models

Impact on data quality solutions and tools

- IT cannot protect anymore the perimeter and quality becomes every consumer’s concern
- Powerful data assessment methods
  - Better understand characteristics of the data with respect to user goals
  - Better support user actions to adjust the content of the data
- Continuous monitoring
  - Leverage analysis of data dependencies and understand critical parts that must be monitored
- Repair the data
  - Support root cause analysis and assistance to where and what to fix
- Key supporting tools: collaborative tagging, versioning, and data lineage

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL

Thank you.

Contact information:

Eric Simon
eric.simon@sap.com

Run Simple

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I INTERNAL