SAP HANA and SAP Vora

Data Management for Enterprises

Stefan Bäuerle, Jonathan Dees, SAP
August, 2017
Agenda

Introduction

SAP HANA
- Overview, Scenarios
- InMemory & Multi-Model

SAP Vora & SAP DataHub
- Motivation
- SAP Vora Scale-Out Architecture and Challenges
- SAP DataHub Architecture and Challenges

Summary
Introduction
SAP is the **world leader in enterprise applications** in terms of software and software-related service revenue. Based on market capitalization, we are the world’s **third largest independent software manufacturer**.

- **335,000+** Customers in more than 180 countries
- **87,100+** Employees in 130+ countries
- **€22.06bn** Total Revenue (IFRS) in FY2016 (preliminary)
- **87%** Of Forbes Global 2000 are SAP customers
- **45 Years** Of history and innovation
- **100+** Innovation and development centers
- **15,000+** SAP partner companies globally
- **125 Mil.** Subscribes in our cloud user base

*Source: http://go.sap.com/corporate/en/company.html*
Key Trends

Digital Transformation
- Customer Experience
- Operational Processes
- Business Models
- Paradigm shift in data processing
- Mobile, cloud, social, big data

Bi-Modal IT
- Managing two coherent styles for work
  - Focus on predictability,
  - Focus on exploration

Big Data & Internet-of-Things
- Big Data for Enterprises
- Combining big data (devices, social) with Enterprise data
Strong Enterprise-level Data Management Background
SAP HANA
SAP HANA: The Modern Data Platform Innovation Journey

Free IT to Focus on Innovation
Leverage ALL Data for Innovation
Discover Deeper Insight to Drive Innovation
Develop Solutions to Power Innovation
Adopt Innovation at The Pace of Your Business
Smart Traffic in the City of Nanjing

**TECHNOLOGY TREND(S)**
- Super Computing
- Smarter World

**BENEFITS**
- Cost reduction
- Customer satisfaction
- Sustainability/CSR

**CUSTOMER SITUATION AND PAIN POINTS**
thousand taxi cabs, 7,000 buses and 1 million private cars running Nanjing, once the capital of China, is one of the top 20 cities in China with a population over 8 million and a GDP of $141.7 billion (with a growth rate of 10% in 2014). It is a highly influential city for all of China.

- The overall traffic volume is immense. Within the city there are about 10,000 g in the city road network.
- Economically, congestions can have a huge impact. For example, Beijing claimed a 12 billion US$ loss due to congestions in 2013.
- Increasing urbanization makes it difficult for city like Nanjing to control traffic.

**NEED FOR CHANGE**
- Plan and optimize Nanjing traffic regulation.
- Proactively manage different traffic and transportation scenarios to provide better services to citizens and help better urban planning.
- Manage next phase of sustained growth with the help of advanced technology from SAP for big data analysis and predictive analytics, mobile computing and cloud-based solutions

**SITUATION AFTER DIGITIZATION**
- Together with SAP, the city of Nanjing developed a next-generation solution, Smart Traffic, which includes the use of sensors and RFID chips to generate continuous data streams about the status of transportation systems across the city.
- Adopting smart traffic control technology to crunch the 20 billion data points captured in the city every year has helped the city to produce actionable information for predictively responding to traffic congestion.
- Smart Traffic solutions, like electronic tolling, traffic management, parking guidance & reservation systems, navigation support, real-time traffic updates and passenger display systems help understand the city’s dynamic travel patterns and determine the impact estimation to eventually mitigate congestions or car accidents.

**TANGIBLE RESULTS**
- Smart Traffic Control enables cities to optimize traffic-light controls and free up additional car lanes during the rush hour to alleviate congestion.
- With SAP’s new approach for zoning -Traffic Analysis Zone (TAZ) the city is able to create one digital map that allows for a holistic view on the as-is situation on the spot and generates recommendations for traffic planning.
- The map can tell exactly how many vehicles arrive or leave a certain city district in the early morning hours. This information is then used to predict the after-work traffic.
- Concrete numbers to be expected.

**SOURCES**
- Smart Cities – How SAP Helps Clear Traffic Jams
- Connected Cars Rev Up For A Revolution
- Nanjing Video
- Smart Traffic Innovations for Cities and Passenger Transit from SAP
Smart Traffic in the City of Nanjing

Platform used to develop the application
SAP HANA Platform

Processor primarily used in the deployment
Intel® Xeon® processor E7

Technical impact of the solution
1. The solution processes over 100 million records of FCD (Floating Car Data) and FDD (Fix Device Data) every day
2. The solution calculates Traffic Performance Index, Road Speed and Road Volume every 5 minutes based on the data volume in point 1
3. City management and 0.8 million citizen uses the result of our solution
Objectives

- Support transformation to become Europe’s leading online fashion platform
- Increase financial processing capacity to handle the demands of the platform business
- Develop a smart data strategy

Why SAP

Financial processes that have run with SAP software for many years

Resolution

- Migrated a contract accounts receivable and payable (FI-CA) solution from SAP that was running on Oracle to the SAP HANA platform
- Became the first SAP customer in the world to run this solution on SAP HANA
- Was selected as a finalist for an SAP HANA Innovation Award

Benefits

- Increased financial processing capacity by an order of magnitude
- Greatly accelerated monthly closing
- Increased employee productivity by eliminating runtime issues
- Improved customer and employee satisfaction as well as attractiveness as an employer
- With the migration to SAP HANA, the system is now ready to handle the new demands. It can conservatively process at least 10 times the current volume

“With SAP HANA, we have substantially accelerated all financial processes from dunning through invoice processing to closing activities. Our staff is now much more confident in their day-to-day work, including interacting with customers.”

Martina Hilzinger, Head of SAP Competence Center, Zalando SE
Creating a Real-Time Supply Chain by breaking down Interenterprise Barriers

Objectives
- Build a real-time supply chain
- Break down information communication barriers between different enterprises, and integrate them to enable real-time information sharing and workflow coordination
- Synchronize collected data in real time via the Internet of Things for production facilities and detection equipment to enable real-time quality control and progress tracking

Why SAP
- Good existing partnership with SAP
- Unmatched performance of the SAP HANA platform

Resolution
- Implemented SAP HANA
- Integrated SAP HANA directly with R database
- Launched the project in three stages: quality management and control, creating an order system and instant-messaging platform, and logistics management

Benefits
- Significant cost reduction
- Greatly improved efficiency
- Greater quality management and control

“SAP HANA is a high-performance computing platform. Without it, there is nothing we can do to extend the supply chain, no matter how superb our ideas and quality-control procedures are.”

Xiangguo Dong, IT Director and Project Lead, Advanced Micro-Fabrication Equipment Inc.

3%–5% Improvement in process capability and process variation
2%–3% Increase in product-test pass rate
30% Reduction in quality cost
0 Production delays caused by quality problems
SAP HANA Platform: The platform for all applications
Simplify, accelerate, innovate

S A P  H A N A  P L A T F O R M

APPLICATION SERVICES
Web Server
JavaScript
Fiori UX
Graphic Modeler
Application Lifecycle Management

PROCESSING SERVICES
Spatial
Graph
JSON
Predictive
Search

INTEGRATION & QUALITY SERVICES
Data Virtualization
ETL & Replication
Data Quality
SAP Vora, Hadoop & Spark Integration
Remote Data Sync

DATABASE SERVICES
Columnar
OLTP + OLAP
Multi-Core & Parallelization
Advanced Compression
Multi-tenancy
Multi-Tier Storage
Data Modeling
Openness
Admin & Security
High Availability & Disaster Recovery

ALL DEVICES
SAP, ISV and Custom Applications

ONE Open Platform
OLTP + OLAP
ONE Copy of the Data
SAP HANA Platform: Database Services

Breakthrough innovations

<table>
<thead>
<tr>
<th>SAP HANA PLATFORM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>APPLICATION SERVICES</strong></td>
</tr>
<tr>
<td>Columnar OLTP+OLAP</td>
</tr>
<tr>
<td>Advanced Compression</td>
</tr>
<tr>
<td>Advanced Compression</td>
</tr>
<tr>
<td>Multi-tenancy</td>
</tr>
<tr>
<td>Multi-Tier Storage</td>
</tr>
</tbody>
</table>

- Turns data into real-time information
- No database tuning required for complex and ad hoc queries
- Run Transactions and Analytics together on one system and one copy of data
- Ready for Cloud, Hybrid, or On-premise deployment
- Not limited by the size of memory
SAP HANA Platform: Database Services
Comprehensive advanced data processing and analytics

<table>
<thead>
<tr>
<th>APPLICATION SERVICES</th>
<th>PROCESSING SERVICES</th>
<th>INTEGRATION &amp; QUALITY SERVICES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial</td>
<td>Graph</td>
<td>Data Business Functions</td>
</tr>
<tr>
<td>JSON</td>
<td>Predictive</td>
<td>Functions</td>
</tr>
<tr>
<td>Machine</td>
<td>Search</td>
<td>Streaming Analytics</td>
</tr>
<tr>
<td>Learning</td>
<td>Text Analytics</td>
<td>Series Data</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Business Functions</td>
</tr>
</tbody>
</table>

**DATABASE SERVICES**

- Run applications with dramatically different datatype characteristics in the same system
- Optimize graph, planning, and rules applications on the same data
- Empower your business via built-in predictive analytics, business functions, and data quality
SAP HANA Platform: Application Services
Web server and database in one system reducing data movements

<table>
<thead>
<tr>
<th>SAP HANA PLATFORM</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPLICATION SERVICES</td>
</tr>
<tr>
<td>Web Server</td>
</tr>
</tbody>
</table>

**SAP HANA PLATFORM**

- Deliver consumer-grade User Experiences for any device, automatically
- Support standards and languages that developers already know how to use
  - Java, JavaScript, C++, Node.js, SQL, JSON, ADO.NET, JDBC, ODBC, OData, HTML5, MDX, XML/A
- Built-in tools to develop, version-control, bundle, transport, and install applications
SAP HANA Platform: Integration Services
Data from any source for a complete view of the business

### SAP HANA PLATFORM

<table>
<thead>
<tr>
<th>APPLICATION SERVICES</th>
<th>PROCESSING SERVICES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>INTEGRATION &amp; QUALITY SERVICES</td>
</tr>
<tr>
<td></td>
<td>Data Virtualization</td>
</tr>
<tr>
<td></td>
<td>ELT &amp; Replication</td>
</tr>
<tr>
<td></td>
<td>Data Quality</td>
</tr>
<tr>
<td></td>
<td>SAP Vora, Hadoop &amp; Spark Integration</td>
</tr>
<tr>
<td></td>
<td>Remote Data Sync</td>
</tr>
</tbody>
</table>

### DATABASE SERVICES

- Access information stored in data silos while keeping the data in place
- Replicate and move any type of data in real-time to the cloud and on-premise when necessary
- Capture and analysis live data streams and route to appropriate storage or dashboard
- Synchronize data between HANA and thousands of remote databases (SQL Anywhere, UltraLite)
- Multiple access points from HANA to BigData: thru SAP Vora, Spark, Hive, HDFS & MapReduce
SAP HANA Demo: Predicting Landslides

Live Earth Data: Analysis & Insight Through Machine Learning
SAP HANA Platform: NoSQL & Multi-Model Data Processing Capabilities

Graph
- A schema-flexible and scalable data store for storage, processing, combination, exploration, and analysis of irregularly structured data with complex and dynamic relationships
- Provides property graph model of Vertices & Edges
- Powerful declarative data query and manipulation language for property graphs
- Imperative API for developing application-specific graph algorithms directly in the database engine
- Cypher for Pattern matching

Multi-Model Data Processing

Document Store
- Enterprise ready Document Store to handle JSON documents natively
- Full ACID support
- Extended processing, e.g. joins between collections, aggregations
- Multi-Model Cross-processing, e.g. Joins
- High performance (even for analytics)
- SQL-like syntax
SAP Vora & SAP DataHub
Big Data? More than a buzzword

Huge Data Volume
Terabytes to multiple Petabyte, low value data

Increasing Data Variety
Diverse sources, many data types: like text, graph, image, custom..

Data quality / data trust
Systematic data errors, format issues, missing structure, missing entries, alignment issues, how can data/source be trusted…

High Data Processing Velocity
High speed streaming from a huge number of different sources
Modern landscapes: an example

Key challenges and questions

- Diverse technologies
- Diverse consumers
- Diverse requirements

- How can huge and ever-growing data volumes be handled?
- Rapidly changing landscape – where is the next opportunity?

Another set of silos is not the answer!
Finalize the vision: flow-based big data management

COLLECT

REFINE

STRUCTURE

MODEL

CONSUME

Data Lake

Distributed Big Data Management

Enterprise Data Warehouse

JOIN, LOCK-UP, FILTER, CLEANSE

SCHEMA, CONNECTIONS, CORRELATION, TYPES

AGGREGATE, FEDERATE, MODEL, TRANSFORM

ANALYZE, EVENTS, TRIGGER

© 2017 SAP SE or an SAP affiliate company. All rights reserved. ǀ CUSTOMER
Toward a cloud-ready and efficient big data processing Architecture

- **Devices**
  - Edge / Device Management
  - IoT Service
  - Message Broker

- **SAP Data Hub**
  - Pipeline Engine
    - Spark
    - Custom Code
    - Rules

- **SPARK**
  - SAP Vora
    - Engine
    - Engine
    - Engine

- **Kubernetes Cluster**
  - Cluster Orchestration

- **Applications**
  - Predictive Maintenance
  - Real-time Demand/Supply Forecast
  - Fraud Detection
  - Brand Sentiment
  - Product Recommendation

- **CONSUME**
  - Personalized Care
    - Interface Collection

- **MODELS**
  - SAP HANA
    - In-memory store
  - External / Cloud Storage

© 2017 SAP SE or an SAP affiliate company. All rights reserved. | CUSTOMER
Manage Big Data with SAP Vora

SAP VORA

Intelligent cluster management
- Partition assignment
- Transaction mgmt
- Node control
- Metadata

In-memory processing engines
- Relational
- Time series
- Documents
- Graph

Disk-based Engine

SAP HANA

In-memory store

Kubernetes Cluster

Data lake or cloud provider
- Spark extensions
  - Hierarchies
  - Unit support
  - Currency conversion
  - Table metadata

Disk-based Engine

Cloud Storage

Model
SQL Extensions for (non)-relational data

One holistic experience => Integrate into one single language (SQL)

**Graph Engine**
create graph mygraph;

select a.name, m.title from actor a, movie m using graph mygraph where m in a.plays_in;

**Document Store**
create collection mydocs;
insert into mydocs { name: "james", age: 42 };
select { age: avg(age), name: name } from mydocs group by name;

**Timeseries**
create table myseries (ts timestamp, i int ) series ( period for timeseries ts start ... end ... equidistant increment by 1 second spline error 1.2 percent);

select median(j) from t1 where period between timestamp ... and timestamp ...;

**Relational Main Memory/Disk Store**
create table mytab( i int ) store on disk;
create table mytab( i int ) store in memory;

select i, count(*) from mytab group by i;
Example: Join Query (1)

```
SELECT p_partkey FROM partsupp JOIN part ON ps_partkey = p_partkey
```

- PARTSUPP partitioned: 4 hosts by partkey
- PART partitioned: 6 hosts with ranges by partkey
Challenges (1): Scalable, Fault Tolerant Algorithms

- Abstraction exchange operator (1:n, n:1, n:m) good enough? (compare: MPI)
- Amdahl’s law: A single non-scalable part prevents overall scalability (order of magnitudes)
  => Make *everything* scalable

- Examples
  - Order by
  - Join
  - Aggregation
  - Window Functions
  - Common Table Expressions, Recursive
  - Table Functions
  - Non-Relational Operators (for Graph, ... )
  - Custom Functions
Data Partitioning - Schemes

- Different partitioning types: hash, range, block, and interval
- Multi-dimensions partitioning
- Consistent hashing for partition to node assignment
- Split partitions for node utilization
- Load pre-partitioned files
Challenges (2): Data Partitioning

- Incremental partition creation (for range and interval)
  - Statistical facts of new data
  - Multi-dimensions partitioning
  - Outliers processing
- Graph partitioning
- Changing partition criteria while staying high available (no table lock)
Toward a cloud-ready and efficient big data processing Architecture

- **Devices**
- **Edge / Device Management**
- **IoT Service**
- **Message Broker**

**Applications**
- Predictive Maintenance
- Real-time Demand/Supply Forecast
- Fraud Detection
- Brand Sentiment
- Product Recommendation

**Personalized Care**

**SPARK**
- Pipeline Engine
  - Spark
  - Custom Code
  - Rules
  - Op1
  - Op 2
  - Op 3
  - Op 4

**SAP Data Hub**

**SAP Vora**
- Processing Engine
  - Engine
  - Engine
  - Engine
  - Engine

**Kubernetes Cluster**

**SAP HANA**
- In-memory store

**External / Cloud Storage**

© 2017 SAP SE or an SAP affiliate company. All rights reserved. | CUSTOMER
SAP Data Hub
Integrate, orchestrate and manage diverse big data infrastructures

SAP Data Hub
- Monitoring
- Orchestration
- Data Management & Preparation
- Data Flow Pipeline Engine

Existing Systems
- SAP HANA
- Data Driven App
- Data Driven App
- Data Driven App

Distributed Data Systems
- Hadoop
- Cloud Storage
- Machine Learning

SAP Cloud Platform
SAP Datahub Pipeline Engine

- Data driven execution
- Operators with In- and Output ports
  - Data transfer via NATS messaging
  - Easy and flexible to define (Python, Scala, Go, C++, …)
  - Run and scale in docker container
- Comes with existing operators and other frameworks (Tensorflow, Spark, R-lang, SAP Vora, SAP HANA, …)
Challenge (3): SAP Pipeline Engine

- Operators stateless
  - + easier scalable and deployment
  - - makes certain implementations more complicated
- Efficient and flexible type specification between operators
  - support runtime + compile time
  - support inheritance
  - multiple languages
  - detect errors early
- Flexible edges without interfering with scalability
Key Take-aways

SAP HANA
- Digital foundation for next-generation applications using latest technology innovations
- Keeping typically enterprise data and other hot data
- Data access layer to additional data like big data handled by e.g. SAP Vora

SAP Vora
- Big Data management infrastructure for the enterprise
- Operating on clusters – providing different engines and storage options

SAP HANA & SAP Vora
- Integrated data management across enterprise Data and Big Data/IoT data

SAP Data Hub
- Combination of core SAP product and platform engagements
- The foundation: SAP HANA Platform + SAP VORA
- Simplifying Enterprise and Big Data processes across distributed data landscapes
Research Areas

Selected list of research topics

- Accelerators
  - General: Use specific hardware technologies to improve SQL processing
  - FPGA
    - Accelerating text search algorithms using FPGAs
    - Leveraging NVMe and FPGAs as a cold store in SAP HANA
  - GPU
    - Re-visiting GPU-based database kernel acceleration with latest advancements in GPU architectures
  - SGX: Evaluation of Flexibility and Scalability of Intel® SGX

- Aging / Data tiering
  - Efficient Query Processing on Hot and Cold Data in In-Memory Databases

- Machine Learning
  - Scalable Distributed Algorithms
  - Efficient Partitioning
  - Dataflow Processing Framework
Thank you.

Contact Information:
Stefan Baeuerle: Stefan.Baeuerle@sap.com
Jonathan Dees Jonathan.Dees@sap.com