SAP HANA –
From Relational OLAP Database to Big Data Infrastructure

Anil K Goel
VP & Chief Architect, SAP HANA Data Platform

WBDB 2015, June 16, 2015
Toronto
SAP Big Data Story

SAP HANA
Big Data Platform

Data Lifecycle Management

Complex Event Processing

Benchmarking

Enterprise Data Consolidation

Data Federation

© 2015 SAP SE or an SAP affiliate company. All rights reserved.
Big Data Dimensions in SAP Products / Technologies
The Native and Open Strategy

✓ Variety
  • In-memory/On-disk Column/Disk stores
  • Full-text capabilities
  • Support for GIS, time series, graph data, …

✓ Velocity
  • SQL-Anywhere for distributed data capturing, local processing, efficient propagation
  • Loading from external data sources via Data Services or Replication Server
  • Event Stream Processor for CEP / low-latency data processing

✓ Volume
  • Multi-TB databases with SAP HANA in-memory DBMS
  • 12PB data warehouse with SAP IQ storage
  • Hadoop for even larger stores
SAP HANA Big Data Platform
Holistic End-to-end Ecosystem

HANA Data Management Platform

- Information Management
- Text
- Search
- Graph
- Geospatial
- Predictive

SAP HANA In-Memory
Instant Results
0.0sec

HANA Dynamic Tiering
Warm Data

HADOOP
HANA SOE
Infinite Storage
Raw / Archive Data

Complex Event Processing / Smart Data Streaming
Bi-directional Replication / Remote Data Sync
Administration | Monitoring | Operations | User Management | Security
SAP HANA Platform for the Complete Solution Space

The Value Dimension

- SAP HANA
- SAP HANA IQ (extended storage)
- Hadoop infrastructure (any distribution)

Data consumption

Value

Data provisioning

Age

HOT

COLD

ARCHIVE
Dynamic Tiering (DT)

- Dynamically partitioned data sets
  - Data Life Cycle / Aging
  - Storage hierarchies
- Unified system → landscape simplification
- Seamless growth and aging of data
- Integrated query processing
- Independent management of storage hierarchies
Dynamic Tiering (DT)

**DDL:**
CREATE TABLE table_name table_definition
USING EXTENDED STORAGE

**Transactions**
- DT Storage participates in distributed transactions [ICDE2013]
  - Internal optimized 2PC like coordination
- Recovery integrated with SAP HANA including point-in-time recovery

**Query Processing**
- Tightly integrated DQP across HANA and DT
- Multiple distributed query processing strategies
  - Remote Scan
  - Semijoin
  - Table Relocation
  - Union Plan
Smart Data Access (SDA)

- Access layer for various remote data sources, e.g. Hadoop, relational and non-relational systems
- Extensible federation layer based
- Expose tables of remote data sources as virtual tables
- Uses capability description for remote data source

- Realizes HANA Open Strategy for Big Data
Smart Data Access (SDA)

Register Remote Date Sources

CREATE REMOTE SOURCE HIVE1
  ADAPTER "hiveodbc" CONFIGURATION 'DSN=hive1'
  WITH CREDENTIAL TYPE 'PASSWORD' USING 'user=dfuser;password=dfpass';

CREATE VIRTUAL TABLE "VIRTUAL_PRODUCT" AT "HIVE1"."dflo"."dflo"."product";

SELECT product_name, brand_name FROM "VIRTUAL_PRODUCT";

Federated Query Processing

- Capability-based framework, e.g. CAP_JOINS : true or CAP_JOINS_OUTER : true
- HANA leverages statistics available via remote sources (e.g. Hive MetaStore) for cost-based optimization
HANA & Hadoop Integration

Hadoop Integration

- SQL on Hadoop via SDA (virtual tables) – Hive or Spark
- Execution of MR-Jobs via HANA (Virtual Functions)
- Access to HDFS (via virtual function)
- Integration on storage & processing
  - Use Hadoop storage (HDFS)
  - Push processing to Hadoop (code to data)
Remote Materialization for Hive

- Batch processing in Hive or MapReduce implies high latencies and response times
- Use remote materialization with configurable data freshness requirements

- **WITH HINT (USE_REMOTE_CACHE)** for SQL query
- Hive extension checks for hint and checks for cached result with configured freshness guarantee
- Non-transactional access for Hadoop-side data
- Cached data is stored in HDFS

- Many enhancements possible
  - On-the-fly materialized view creation
  - Remote and / or local
WBDB 2014: SAP Standard Application Benchmarks

Goals

- Analyze and optimize performance of SAP components and business scenarios
- Compare the performance of computer systems from different vendors
- Create an open competition between different vendors by providing the possibility to publish results
- Provide input for initial sizing
  - Throughput numbers are defined in business application terms, e.g. “fully processed order line items per hour”
  - Business throughput is mapped onto the resource consumption of the most prominent hardware components, incl. CPU (SAPS) and memory

Method is monitored and approved by the SAP Benchmark Council

Business-relevant results

- Free from artifacts, customers can rely on the results
- Hardware and software combination must be available for customers
- Only configurations that can be used in production environments are permitted
WBDB 2014: BW-EML Standard Application Benchmark Facts

- **BW-EML Benchmark Workload:**
  Mixture of multiuser query load accessing all 10 available InfoProviders

- **Simultaneous Delta Loads:**
  During high load reporting activity, static operational data is extended with delta data every 5 minutes with a total of 1/1000 the original data set into all InfoProviders

- **High Load Phase:**
  High load phase is minimum 1 hour.

- **Benchmark Key Performance Indicator:**
  The key figure of this benchmark is the number of ad-hoc navigation steps/hour.

- **Multiuser query load:**
  Each benchmark user runs all queries / navigation steps sequentially in a configurable number of loops

- **Navigation Steps per Loop:**
  Total number of navigation steps per loop: 40
SAP Big Data Benchmarking Considerations

Data types and life cycle across the platform

- Multi-store: transactional (in-memory/on-disk), MPP, Hadoop, …
  - Dynamic Tiering / Aging / Data Life Cycle
  - Impedance mismatch
- Streaming
- Multi-data
- Multi-workload
  - Analytical: predictive, machine learning, typical Big Data, …
  - Transactional: business data, systems of record, low latency, …
  - Hybrid: Distributed data center

(Industrial) IoT as a special use case

- Machine data
- Series data
- Edge data processing
SAP Big Data Benchmarking Considerations

Data Generation
- Customer specific
- Realistic
- Scalable
- “Dexter” @SAP
- “Myriad-Oligos” @TU Berlin

Diversity
- Applications
- Systems
- Languages
SAP HANA Big Data Platform
Summary

SAP HANA Native Big Data Capabilities

Optimal Integration of Non-SAP Big Data Technologies into SAP Business Landscapes
Application & Analytics Capabilities

Hadoop / Others as Strategic Big Data Technologies for SAP Extend With Our Own Engines and Analytics (HANA SOE)
Anil K Goel
anil.goel@sap.com