SAP Sybase IQ 16 In-Database Analytics Option Technical Overview

Andrew Neugebauer/Analytics Product Management
March 08, 2013

Agenda

What's Happening in the Marketplace Today

SAP Sybase IQ Product Success

SAP Sybase IQ 16

In-Database Analytics Landscape

Base Product In-Database Analytics Features

In-Database Analytics Option Features

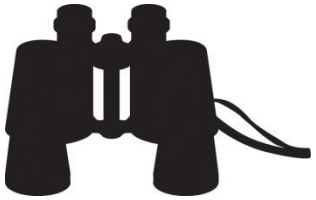
Summary

Marketplace Today

What's Happening in the Marketplace…

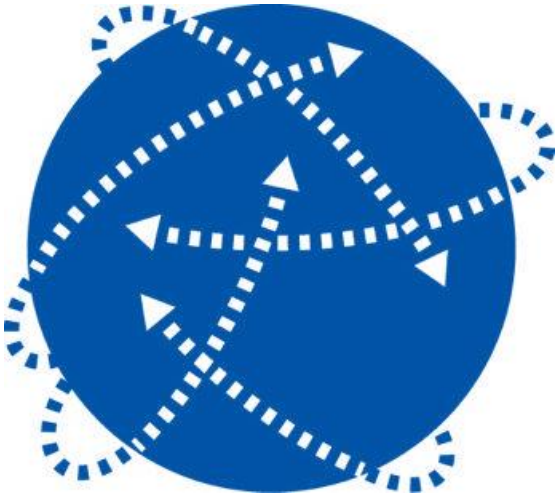
Exploding Data Volumes | **The Need for Speed** | **Rising IT Cost and Complexity**

Challenges customer face today?



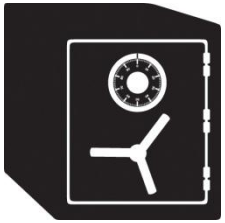
Lost revenues due to lack of insight

High Costs & Complexities



Slow Performance

Data Management Challenges

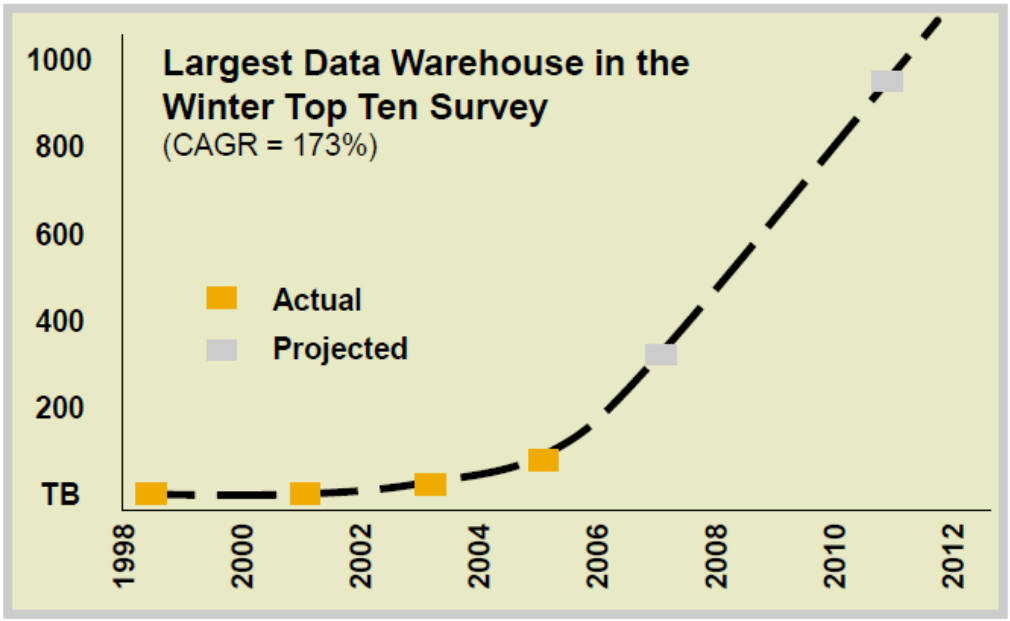


Security

SAP Sybase IQ 16 Motivators

"Petabyte is the new Terabyte" - Forbes

The data explosion continues: Data volumes in analytics environments are growing exponentially…



Source: *WinterCorp*

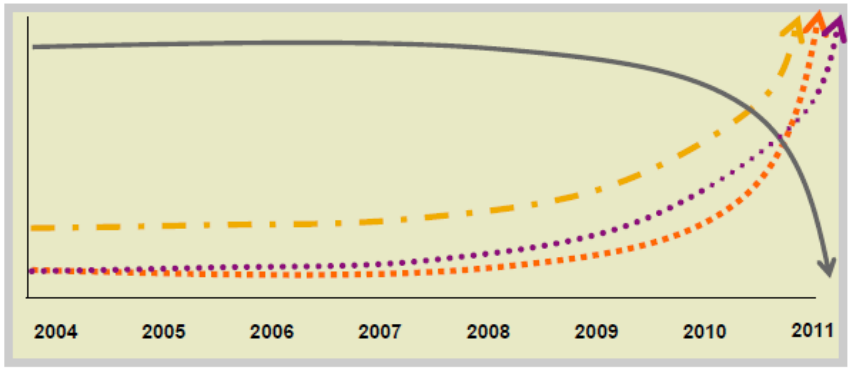
Meanwhile, what trends are you seeing…

Number of business decisions supported by data?

Number of users accessing analytics environments?

Required response time?

Resources required to maintain performance?



Product success

SAP Sybase IQ: Market Leader for Extreme-Scale EDW and Analytics

High performance analytics server

- Columnar RDBMS (stores data in columns- versus rows)
- Optimized for managing and accessing massive amounts of data for analytics (versus transactions)

Accelerates analytics and reporting

- Up to 1000-times faster than traditional transactional databases
- Handles structured and unstructured data
- High compression and low TCO
- Highly scalable grid architecture

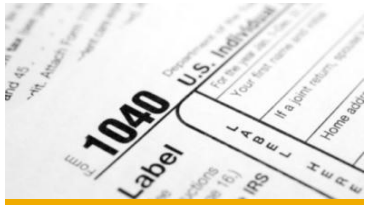
SAP Sybase IQ Facts

- 2200+ customers with over 4500+ installations worldwide
- Used by twice as many companies as the next leading provider
- Patented data compression dramatically reduces data storage requirement; cuts TCO
- Only column-based solution to support full text search, in-database analytics, and federated analytics
- 96%+ customer satisfaction rates
- Leader, 2013 Gartner Magic Quadrant for Data Warehouse DBMS

SAP Sybase IQ big data analytics

Pervasive across data intensive industries worldwide

Manage and analyze statistical measures for the entire nation of Canada

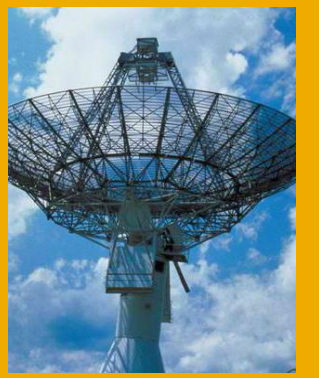


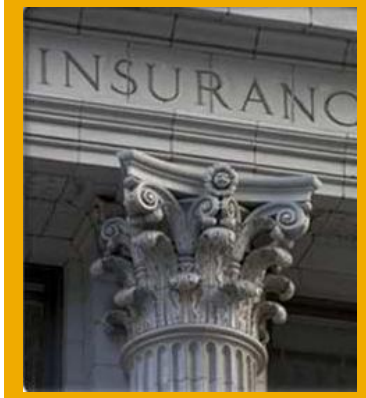
Analyze ALL Federal tax returns in the US

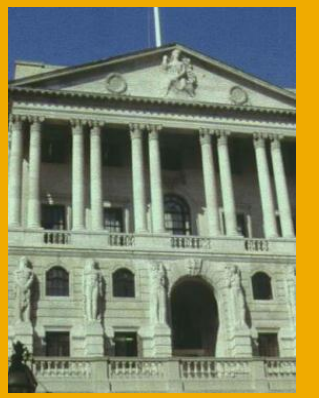
Analyze complex models in more than 200 financial institutions worldwide

Store and analyze massive amounts of industry segment data in 30 of the largest information providers in the world, including Transunion, Nielsen and Axiom

Stands out as the leading enterprise data warehouse among the largest banks, insurance agencies, and telecom operators worldwide







SAP Sybase IQ 16

Solution Overview – SAP Sybase IQ 16

SAP Sybase IQ transforms the way companies compete and win through actionable intelligence delivered at the speed of business to more people and processes.

Value of SAP Sybase IQ 16

1. **Exploits the value of Big Data**
2. **Transforms businesses through deeper insights**
3. **Extends the power of analytics across the entire enterprise**

In-Database Analytics Landscape

Traditional complex analytics

Data to logic

Fetch data from database

Create datasets for analytical packages

Time consuming process
Could run into memory constraints with large data sets

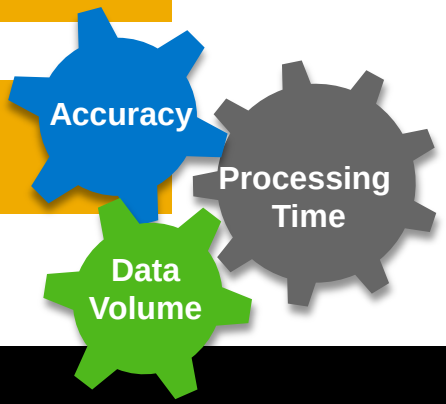
Analyze data using statistical functions on proprietary platforms

Proprietary platforms make it very difficult to embed in applications

Store results from datasets back into database

Another time consuming process which could slow down the delivery of results to end-users

Generate reports



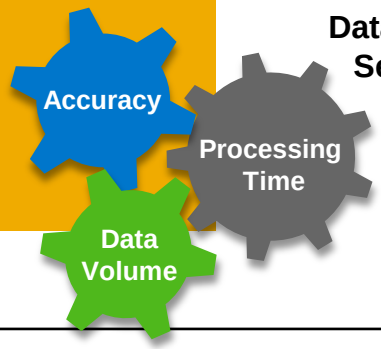
Compromise on at least one key area

In-database analytics with SAP Sybase IQ

Logic to the data

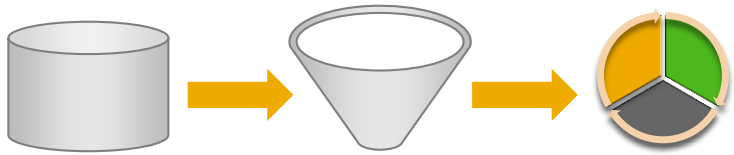
Logic to data: fast + efficient

- Data NEVER leaves database until results are materialized
- Analytics code/models SHAREABLE and allow AD-HOC analysis
- Analytics code/models are applicable to the LATEST data set
- PERFORMANCE and SCALABILITY improvements are visible
- Average BI specialist able to code In-database analytical models



Data to logic: slow + clumsy

Traditional constraints of data analysis must be removed



Database Server

Logic/Filtering Applied Outside the Database Server

Visualization

Vertical	Use Case	Analytical Algorithms
Banking	Up sell / cross sell to improve revenue potential	Clustering — K-means, hierarchical O order
Marketing services	List processing services to improve marketing campaigns	Regressions — linear and logistic
Others	…..	…….

SAP Sybase IQ

Three pronged strategy for in-database analytics

Built-In Native

- Full text search
- OLAP
- Numeric, string, math

UDF Plug-In Native

- Text analytics
- Data quality

UDF Plug-In External

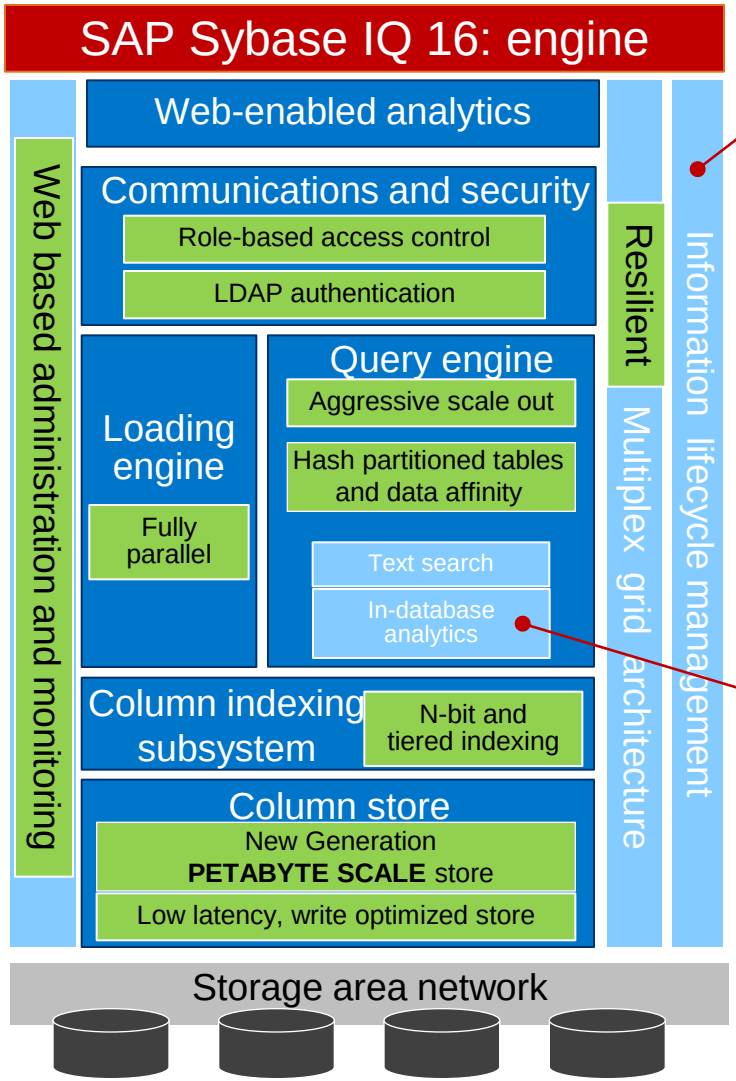
- Descriptive statistics, data mining, simulation
- PMML
- GeoSpatial
- Financial
- Text mining
- Multimedia

- Built-in specialized data types / indexes: Text, Spatial, IP, UDT, UDI, Semantic
- Multi-lingual support: C/C++, JAVA, PERL, Python, CLR, Ruby
- Multi-modal execution: In-Process/Out-Of-Process, SMP / MPP
- Multi-class support: Scalar, Aggregate, Table, Table Parameterized

Base Product In-Database Analytics Features

SAP Sybase IQ 16

In-database analytics



Manage data through its existence in the data warehouse

Super-fast in-kernel quantitative analytics

Library of statistical built-ins

Support for built-ins

- DATEFLOOR
- DATEROUND
- DATECEILING
- STDDEV_POP
- STD_SAMP
- ……..

Support for operators

- +-*/ ARITHMETIC FUNCTIONS OVER UNARY
- +-*/ ARITHMETIC FUNCTIONS OVER N-ARY
- UNION (Concatenation) OVER MULTIPLE
- INTERPOLATION
- COUNT
- COUNT DISTINCT
- …….

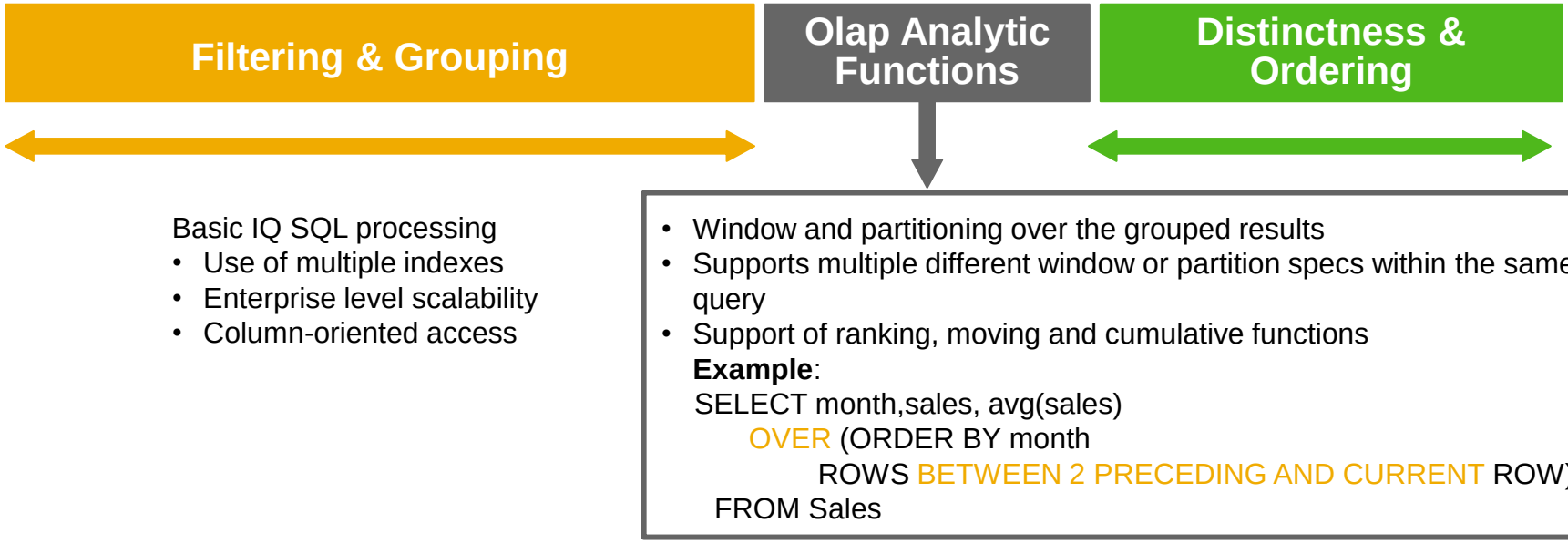
Library of ANSI SQL OLAP built-ins

Support standard ANSI SQL OLAP aggregates

- FIRST_VALUE, LAST_VALUE
- CORR, COVAR_POP, COVAR_SAMP, CUME_DIST
- REGR_AVGX, REGR_AVGY, REGR_COUNT,
- REGR_INTERCEPT, REGR_R2, REGR_SLOPE
- REGR_SXX, REGR_SXY, REGR_SYY

Add Sybase extensions for financial analytics

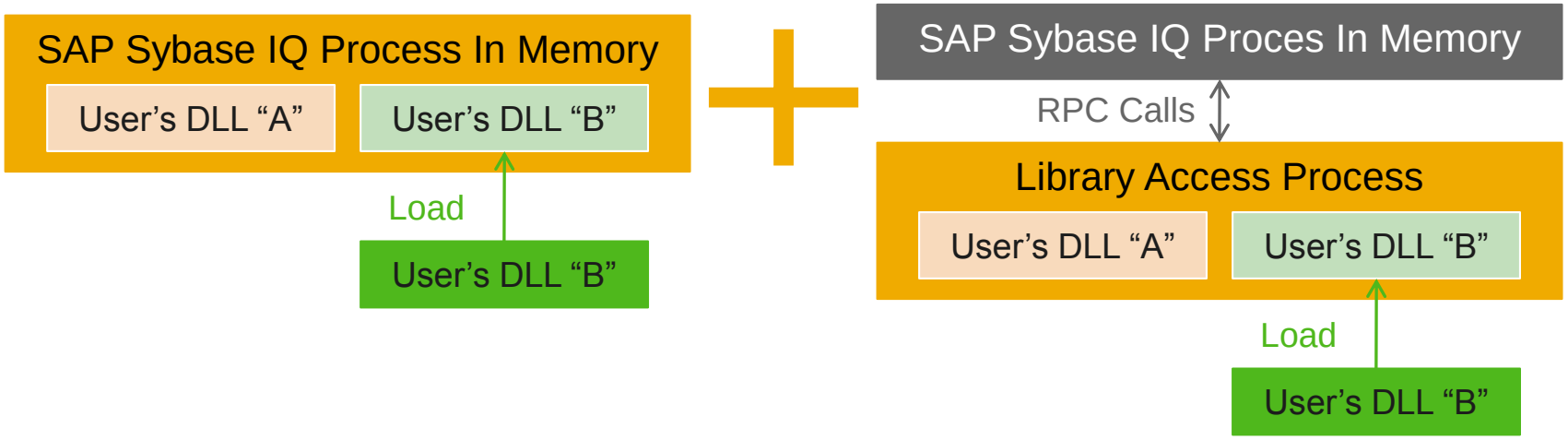
- AW_AVG (arithmetically weighted avg)
- EW_AVG (exponential weighted moving avg)



In-Database Analytics Option Features

SAP Sybase IQ in-database analytics

In-process and out-process



In-database + in-process

- In-process dynamically loaded shared libraries
- Highest possible performance
- Incurs security risks, but manageable via privileges
- Incurs robustness risks, but manageable via multiplex
- C/C++

In-database + out-process

- Out of process shared library
- Lower security risks
- Lower robustness risks
- Lower performance than in-process but better than out of database
- JAVA

Java user-defined functions

Feature

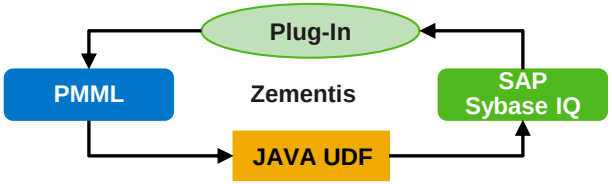
JAVA User Defined Function offers a new in-database analytics API

Characteristics

- External algorithms written as JAVA fns, plugged into SAP Sybase IQ
- JAVA fns via SQL: runs in-database, much faster than client side
- JAVA fns run protected/fault tolerant (in separate process)
- Supports scalar and table outputs
- Supports all data types

Big Data Use Cases

- Ideal for ISV or custom data mining libraries for **Healthcare, eCommerce, Public Sector.**
  Apps include:
  - ISV partner ***Zementis*** built a plug-in for PMML (Predictive Modeling Markup Language) models
  - Validates PMML from SAS, R,..
  - Translates PMML to JAVA UDFs
  - JAVA UDFs called from SQL



C++ user-defined functions

Feature

C++ User Defined Function offers a new In-Database Analytics API

Certified ISV C++ functions can be run on Reader or Writer Nodes

Customer-built C++ functions can be run on Reader Nodes

C++ UDFs

Characteristics

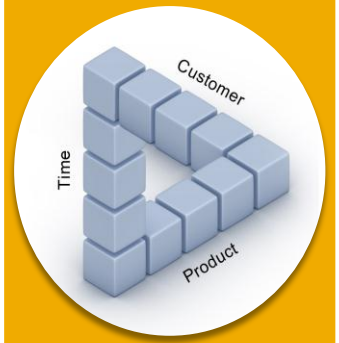
- External algorithms written as C++ functions, plugged into SAP Sybase IQ
- C++ functions via SQL: runs in-database, much faster than client side and JAVA
- C++ functions run SAP Sybase IQ in-process space and parallelized
- Supports scalar, aggregate and table outputs
- Table parameterized UDF API for bulk data exchange (inputs /outputs)
  - Massively parallelizable and distributed
  - Supports Native Map Reduce
- Supports all data types

Big Data Use Cases

- Ideal for standard ISV or custom data mining libraries for **Marketing Services, Financial, Telco.**
- ISV partner FuzzyLogix library of functions:
  - Data Mining
  - Distributions
  - Univariate
  - Multivariate
  - Statistical
  - Mathematical

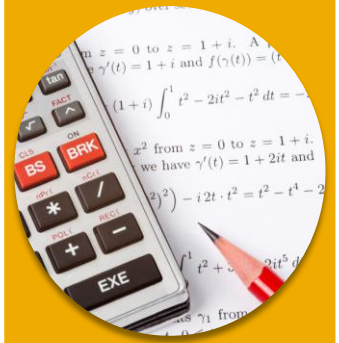
In-database analytics in SAP Sybase IQ

Pre-built functions available natively and through partners



OLAP

- Windowing
- Ranking
- Cubes
- Roll ups
- Correlation
- Covariance
- Weighted average



Mathematical

- Basic math
- Matrix algebra
- Gamma and beta functions
- Area under curve
- Interpolation methods

Statistical

- Descriptive statistics
- Distance measures
- Hypothesis testing
- Cross tabulation
- Anova

Univariate Distributions

- Monte Carlo simulation
- 30 univariate distributions available

Data Mining

- Linear regression
- Logistic regression
- Principal component analysis (PCA)
- Cluster analysis — 5 models available

In-database analytics in SAP Sybase IQ

Custom function development APIs

Several different forms of C++ and JAVA UDF APIs for building custom in-database analytics, each valid at different locations within queries

1. {Scalar} to {Scalar functions} e.g. sin, cosine, …

What: UDF inputs scalar values, outputs a scalar values (C++/JAVA)	**Example:** SELECT my_sin(t.x) FROM t WHERE t.z = 2	**How:** my_plus () scalar UDF computes scalar functions based on single argument passed from SQL

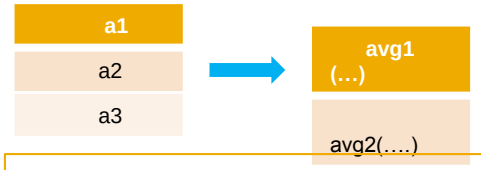
2. {Scalar set} to {Scalar functions} e.g. max, min, …

What: UDF inputs set of scalar values, outputs a relational set of values as a table (C++)	**Example:** SELECT u.r1 FROM my_row_avg (10, 12, 14) WHERE u.r1 >=10	**How:** my_row_avg () UDF computes aggregate values based on a set of input arguments passed from SQL

In-database analytics in SAP Sybase IQ

Custom function development APIs

3. {Scalar set} to {Scalar set} e.g. OLAP windows, …



What: UDF inputs scalar values, outputs a relational set of values as a table (C++)

Example: SELECT month,sales, **my_avg**(sales) OVER (ORDER BY month ROWS BETWEEN 2 PRECEDING AND CURRENT ROW) FROM Sales

How: my_avg() scalar UDF returns 10 rows together that can be joined with other tables

4. {Scalar set} to {Tables} e.g. use attributes from UDF, …

What: UDF inputs and outputs relational set of values as tables; takes scalars (C++, JAVA)

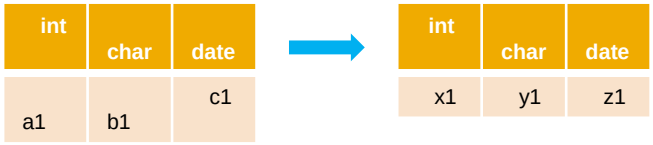
Example: my_sum processed as SELECT my_table.count FROM **my_sum**(TABLE (SELECT dept_id FROM employees)) AS my_table;

How: my_sum table UDF returns a table with attributes based on inputs from a set of values to/from SQL. Output table attributes in the query

In-database analytics in SAP Sybase IQ

Custom function development APIs

5. {Scalar set, Tables} to {Tables} e.g. MapReduce, …



What: Tbl Param UDFs (TPF) to write data intensive/parallel applications e.g. MapReduce (C++)

Example: select new_qts.* from misg_qts (TABLE (SELECT * FROM iq_qts_tbl.qts WHERE ticker IN ('SY', 'TDC', 'ORCL')) OVER (**PARTITION BY** qts.ticker **ORDER BY** qts.trade_time ASC))) AS new_qts

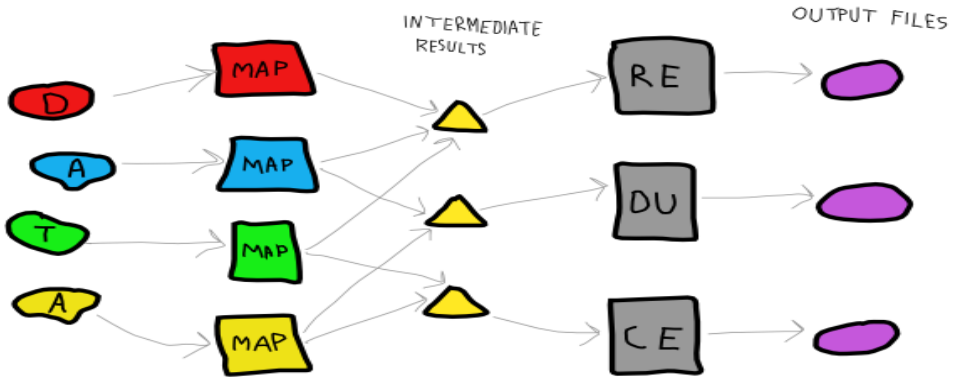
How: SQL caller partitions data into disjoint sets and feeds TPF; TPF outputs table that can be used in the SQL query

In-database analytics in SAP Sybase IQ

MapReduce introduction

- A framework for distributed computing on large data sets on clusters of computers. Processing can occur on data from **file systems** or **DBMS**.

- The framework is based on **map** (distribute work) and **reduce** (collate & output results) steps. Map and Reduce functions can be written and accessed via many languages.



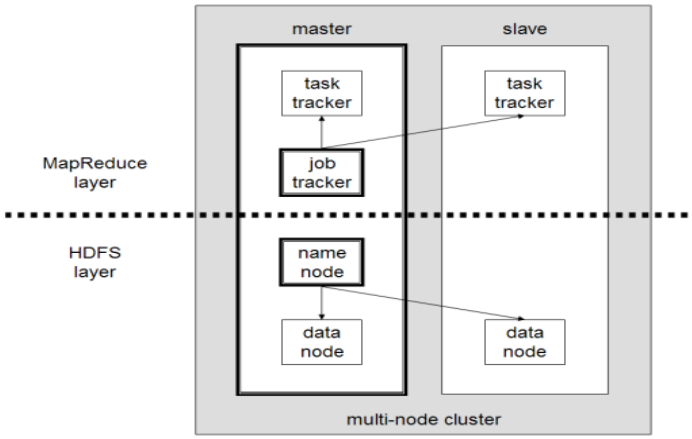
- MapReduce technique is widely used for pre-processing web logs, text data, graph data etc.

Examples: find web links for search, find patterns in social network graphs, etc.

In-database analytics in SAP Sybase IQ

MapReduce in Hadoop

- Open source software framework for MapReduce on large data sets in a distributed file system
  - Storage Layer: Hadoop Distributed File System (HDFS) – stores data in files (schema-free)
  - Application Programming Interface: MapReduce framework



- Commercial support available from – Cloudera, HortonWorks, IBM, EMC/Greenplum
- SWOT
  - Strengths: scalable, cost effective, fault tolerant
  - Weaknesses: batch oriented, weak support for BI tools, tedious programming interface
  - Opportunities: coexistence in large installations can be lucrative (ComScore)
  - Threats: repurposed over time to become a general purpose DW or Analytics Platform

In-database analytics in SAP Sybase IQ

MapReduce in Hadoop vs. SAP Sybase IQ

MapReduce Characteristics

I. Map/Reduce functions consume/produce data sets in bulk

II. Map/Reduce functions executed as parallel jobs i.e. ***n*** Map functions, ***m*** Reduce functions execute independently in parallel

III. Map/Reduce functions work on disjoint data sets i.e. ***n*** Map functions consume ***n*** disjoint data sets producing ***m*** disjoint data sets consumed by ***m*** Reduce functions

IV. Several levels of nested Map/Reduce functions possible — multi-level tree execution

V. Map/Reduce functions triggered by master node but are fault tolerant for worker units i.e. if a worker unit for a fn fails — picked up, completed by another worker unit

VI. Map/Reduce fn libraries written in many languages including popular C++

SAP Sybase IQ TPF for MapReduce

I. TPFs consume/produce data sets in bulk

II. TPFs run in parallel based on ***n*** paths for Map, ***m*** paths for Reduce — decided by SAP Sybase IQ optimizer

III. TPFs are fed with ***n*** disjoint data sets (Map TPFs) that produce ***m*** disjoint data sets (Reduce TPFs) via data partitions specified as part of SQL query

IV. TPFs can be arbitrarily nested to multiple levels via sub-queries

V. TPFs are initiated as parallel work units by a leader node in SAP Sybase IQ PlexQ™ are fault tolerant — if worker node fails, leader picks up, completes work unit

VI. TPFs currently available in popular, performance efficient C++

In-database analytics in SAP Sybase IQ

Pushdown with native MapReduce

For stocks in enterprise software sector, find max relative strength of a stock for a trading day*

Key (k1)	Value (v1)		
30-min interval time	**Ticker Symbol**	**TickValue Day 1**	**TickValue Day 2**
9:30 am	SAP	51	52.4
9:30 am	ORCL	31	28.2
9:30 am	TDC	22	21.3
10:00 am	SAP	50.9	53.1
10:00 am	TDC	21.8	20.9
10:00 am	ORCL	29.4	27.1
…..	ORCL	……	…..

Map Fn

Key (k2)	Value (v2)	
Ticker Symbol	**30-min interval time**	**Weighted variance = (A given stock's variance / Average Variance across All "N" stocks)**
SAP	9:30 am	+1.4 / (SUM (+1.4-2.8-0.7….)/"N" stocks)
SAP	10:00 am	+2.2 / (SUM (+2.2-2.3-1.1 ….)/"N" stocks)
SAP	……	……
ORCL	9:30 am	-2.8 / (SUM (+1.4-2.8-0.7….)/"N" stocks)
ORCL	10:00 am	-2.3 / (SUM (+2.2-2.3-1.1 ….)/"N" stocks)
ORCL	…….	…..
TDC	9:30 am	-0.7 / (SUM (+1.4-2.8-0.7….)/"N" stocks)
TDC	10:00 am	-1.1/ (SUM (+2.2-2.3-1.1 ….)/"N" stocks)
TDC	…..	……

Reduce Fn

	Value (v3)
Ticker Symbol	**Max Absolute Weighted Variance (v3)**
SAP	Max (ABS(9:30 Wt Var), ABS(10:00 Wt Var), …..)
ORCL	Max (ABS(9:30 Wt Var), ABS(10:00 Wt Var), …..)
TDC	Max (ABS(9:30 Wt Var), ABS(10:00 Wt Var), …..)

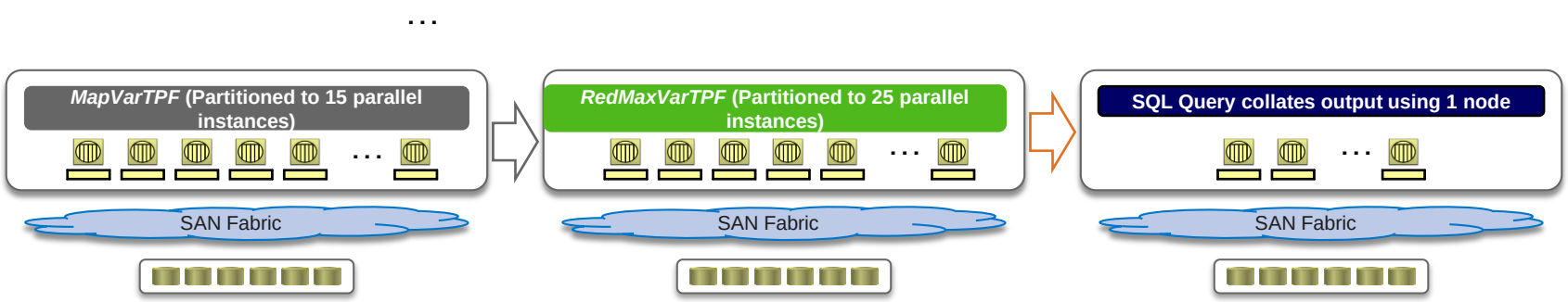
*Calculate max variance for the day by comparing each 30-min interval tick values across two days: the trading day & the day before, weighted by average variance of all stocks for each 30-min interval

In-database analytics in SAP Sybase IQ

Pushdown with native MapReduce

For stocks in enterprise software sector, find max relative strength of a stock for a trading day

- **Map TPF declaration:** CREATE PROCEDURE ***MapVarTPF*** (IN XY TABLE (a1 char, a2 datetime, a3 float, a4 float)
  RESULT SET YZ (b1 char, b2 datetime, b3 float)
- **Reduce TPF declaration:** CREATE PROCEDURE ***RedMaxVarTPF*** (IN XY TABLE (a1 char, a2 datetime, a3 float)
  RESULT SET YZ (b1 char, b2 float)
- **SQL Query:** SELECT RedMaxVarTPF.**TickSymb**, RedMaxVarTPF.**MaxVar**,
  FROM ***RedMaxVarTPF*** (TABLE (SELECT MapVarTPF.TickSymb, MapVarTPF.30MinIntTime, MapVarTPF.Var
  FROM ***MapVarTPF*** (TABLE (SELECT TickDataTab.TickSymb, TickDataTab.30MinIntTime,
  TickDataTab.30MinValDay1, TickDataTab.30MinValDay2)
  OVER (**PARTITION BY** TickDataTab.30MinInt)))
  OVER (**PARTITION BY** MapVarTPF.TickSymb))
  ORDER BY RedMaxVarTPF.TickSymb
- **Native MapReduce parallel execution workflow:**



- **Native MapReduce with unstructured data:** MapReduce using SAP Sybase IQ TPF can easily be applied to unstructured data also e.g. text, multi-media, … stored in Sybase IQ column store or to unstructured data brought into SAP Sybase IQ during execution time from external files

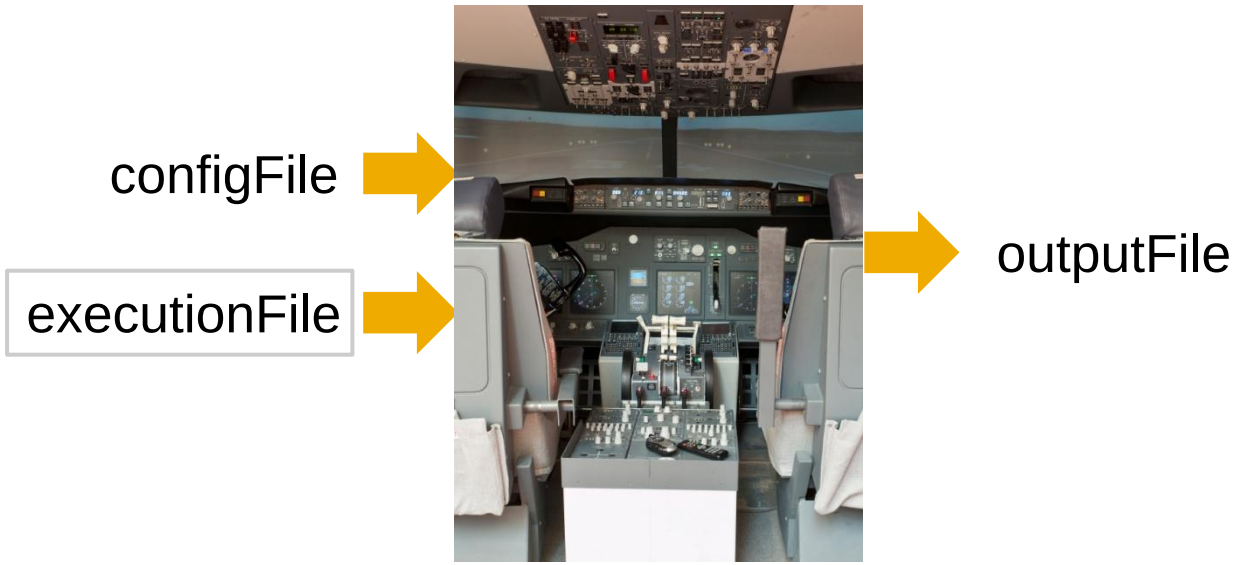
In-database analytics development

UDF simulator

What: Productivity tool for developers to test/debug UDFs w/o a running IQ server; applies to Table UDF/TPF

Example: **udfSim** –c *<configFile>* -e *<executionFile>* [-t][-d[<level:0=least verbose>]] [-parse-only] [-ROOT=*resultId*] OR udfSim *@paramFile*

How: Reads from *configFile* for syntax, from *executionFIle* for execution plan, builds in-memory structures, outputs results in file



In-database analytics in action

Large marketing services firm

Challenge

- Accelerate the speed of model development for their clients

Models used

- Sparse Matrix Calculations, Correlation, Euclidean Distance

Results

- Cut model run time for development models from 20 hours to 30 minutes
- Estimated ROI is $5M in 12 months

Summary

In-database analytics offers

- Native and partner functions
- In-database and out-of-database execution
- In-process with five types of interfaces
- TPF interface well suited for MapReduce

Learn more

- Visit: http://www.sap.com/iq

Thank you

© 2013 SAP AG. All rights reserved.