Apache Trafodion™ (incubating)
Enterprise-Class Transactional SQL-on-Hadoop DBMS

trafodion.apache.org
Rohit Jain is the CTO at Esgyn working on Apache Trafodion™, currently in incubation. Trafodion is a transactional SQL-on-HBase RDBMS. Rohit worked for Tandem, Compaq, and Hewlett-Packard for the last 28 of his 39 years in application and database development. He has worked as an application developer, solutions architect, consultant, software engineer, database architect, development and QA manager, Product Manager, and Chief Technologist. His experience spans Online Transaction Processing, Operational Data Stores, Data Marts, Enterprise Data Warehouses, Business Intelligence, and Advanced Analytics, on distributed massively parallel systems.
Apache Trafodion™

Open source project to develop operational SQL-on-Hadoop database engine

Apache HBase™ + Transactional SQL

Rides the unstoppable Apache Hadoop™ wave!
Transforms how companies store, process, and share big data
Affordable performance, elastic scalability, availability

Open source project - downloadable for free
Apache Trafodion™ is currently undergoing Incubation at the Apache Software Foundation
Eliminates vendor lock-in and licensing fees
Leverages community development resources and speed

Schema flexibility and multi-structured data
Capturing and storing all data for all business functions

Full-function ANSI SQL with JDBC/ODBC access
Leverages existing SQL skills, tools, & apps for productivity

Distributed ACID transaction protection
Data consistency across multiple rows, tables, SQL statements

Targeted for operational workloads!
Optimized for real-time transaction processing applications, operational reporting, and Operational Data Stores (ODS), needing sub-second response times at high levels of concurrency

Data federation: Trafodion/HBase/Hive tables
Enables multiple data model deployment with schema flexibility
## Types of workloads

### OLTP
- Mostly transactional
- Sub-second response
- Customer experience
- Large update volume
- High concurrency
- Scales linearly
- Normalized data model
- Custom applications or 3rd party solutions
- Mostly SMP; MPP for web-scale
- Keyed updates/queries

### ODS
- Can be transactional
- Sub-second to seconds
- Customer experience or Business internal
- Batch to streaming feeds from OLTP
- Low update volume
- Low concurrency if internal, high otherwise
- Near linear scale
- Historical data
- Normalized data model
- Custom apps / 3rd party
- Keyed queries

### BI
- Non-transactional
- Seconds to minutes
- Business internal
- Batch to streaming feeds from OLTP/ODS
- No direct updates
- Low to high concurrency
- Less linear in scale
- Historical data
- Dimension data model
- BI tools – reporting & dashboards
- Ad hoc & scheduled queries and large extracts

### Analytics
- Non-transactional
- Minutes to hours
- Business internal
- Batch/aggregates from BI
- No direct updates
- Low concurrency
- Complex queries, non-linear scale
- Historical & big data
- Columnar store
- Analytics in database
- Analytical tools
- Ad hoc queries

---

**Essential to operate the business**

**To improve performance of the company**
Operational Workloads come to Apache Hadoop™

Operational  Business Intelligence  Analytics

- Enterprise Resource Planning
- Customer Relationship Management
- Manufacturing Resource Planning
- Shared Cache
- Supply Chain Management
- Financial Resource Management
- Human Resource Management
- Data movement & duplication

- Operational Workloads
- Business Intelligence Workloads
- Analytics Workloads

- Column store for fast analytics
- Data movement & duplication

- Transform
- Modernize
- Offload

- Complement
- Offload

- Apache Hadoop
- SQL Server
- Teradata
- Oracle
- Netezza
- Greenplum
- MySQL
- PostgreSQL

- Open Source
- Apache Hadoop
- Apache Tez
- Apache Impala
- Apache Drill
- Apache Spark
- Apache Trafodion
- Apache Hive
- JSON
- Avro
- Orc
- Parquet
Banking

NonStop Mission Critical OLTP system

IBM Mainframe

• Transform → Enrich data
• Modernize → Enhance UX
• Offload
  o Online access
  o Statements
  o Transactional

Multiple years of transactions & statements
- Transform
- Modernize
- Offload

Trafodion for transactions to operational reporting

For closed loop analytics

Semi-structured data

Unstructured data

Intelligent Network (IN), Home Location Register (HLR), Mobile Switching Center (MSC), SMS Center (SMSC), and network elements for other value added services like Push-to-talk (PTT), Ring Back Tone (RBT)
Online Retail …

- Integration of structured, semi-structured, and unstructured support
- Integration of operational, historical, & external (Big) data along common master data for better insights

Structured

<table>
<thead>
<tr>
<th>Item id</th>
<th>Description</th>
<th>Cost</th>
<th>Price</th>
</tr>
</thead>
</table>

Semi-structured

<table>
<thead>
<tr>
<th>TV</th>
<th>Type</th>
<th>Display Size</th>
<th>Resolution</th>
<th>Brand</th>
<th>Model</th>
<th>3D</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book</td>
<td>ISBN</td>
<td>Author</td>
<td>Publish Date</td>
<td>Format</td>
<td>Dept</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unstructured

- Image
- Review

Open distributed HDFS structures

HBase & Hive

Capture data directly into open file structures

Free at last!

Access to Spark

Accessible for reporting & analytics with no latency

SELECT all TVs WHERE Price > 2000 and Type = 'Plasma' and Display Size > '50' and customer sentiment is very positive
Online Retail ...

Asset Management
- Create album
- Upload / Import pictures into album
- Create a project / photo book
- Share album / project with family / friends

Shopping
- Print Calendars, Cards, …
- Order prints, mugs, linen, jewelry, cases, covers, cards, teddy bears, …

Versus RDBMS & NoSQL
- High concurrency low latency workloads
- Limitless elastic scale
- Very low TCO
Online Retail …

Versus RDBMS & NoSQL
- Rich ANSI SQL RDBMS features
- Full ACID transactional support
- Integration of structured, semi-structured, & unstructured data

Online Retail Workloads
- Backend operational workloads: order tracking, supply chain, inventory control, …
- OLTP: Upload pictures, Create album, Tag pictures, Share pictures, Share pictures
- ODS: Search for pictures

OLTP
- Upload pictures
  - Pictures loaded into HDFS by app
  - BEGIN WORK
  - INSERT list of pictures uploaded into Trafodion table PIC
  - (cust_id, album_id, pic_id, pic_date, …)
  - INSERT picture attributes from camera into HBase table PIC_ATTR as col-value pairs for each of the pictures using pic_id
  - END WORK

OLTP
- Create album
  - BEGIN WORK
  - INSERT into Trafodion table ALBUM
  - (cust_id, album_id, album_name, …)
  - END WORK

OLTP
- Tag pictures
  - BEGIN WORK
  - INSERT custom tags for each tagged picture into HBase table PIC_ATTR as col-value pairs
  - END WORK

OLTP
- Share pictures
  - INSERT into Trafodion table REL
  - (cust_id, rel_with_cust_id, rel-type, …)
  - BEGIN WORK
  - INSERT list of pictures shared into Trafodion table SHARED_PIC
  - (pic_id, rel_with_cust_id)
  - END WORK

OLTP
- Order photo mug & jewelry
  - BEGIN WORK
  - INSERT into ORDER
  - (cust_id, order_no, order_date, order_total, …)
  - INSERT into ORDER_DETAIL
  - all items that are part of the order
  - (cust_id, order_no, item_id, pic_id, qty, amt, …)
  - END WORK

ODS
- Search for pictures
  - SELECT pictures taken with my “Sony DSC-RX100M2” camera in the last 6 months from my “Travel” album with a tag “Emma” on it.
Online Retail

Web app

Using model & customer score / attributes, and recent purchase history to make recommendations.

Rohit, consider a blanket for your granddaughter at 50% off with her image imprinted on it.

Versus RDBMS & NoSQL
- Data captured in an open file system with open APIs
- Is available with no latency for reporting & analysis
- Via a huge open source & proprietary Hadoop eco-system

OLTP

Analytics
- BI reporting
  - Sales growth by product, region, demo
  - Growth in customers, pictures, storage, …
  - Growth in sharing
  - …

- Analytics
  - Items bought together – market basket analysis
  - Promotion success / customer classification
  - …

BI reporting

Spark

OLTP

Trafodion

Reporting & Analytics via Spark
Why Apache Trafodion™?
Ingredients for a world class RDBMS

1. **Time, Money, and Talent**
   - 20+ years of investment
   - $300+ million invested since 2006
   - Database developers grew up on
     - Shared nothing Massively Parallel Architecture
     - With a single system image across clusters
   - 300+ years of database experience
     - On building OLTP and BI engines

ANSI and non-ANSI functionality supported, performance, scalability, concurrency, throughput, stability, high availability, transactional, and myriad of other capabilities across a multitude of workloads
Why Apache Trafodion™?
Ingredients for a world class RDBMS

2. World Class Optimizer

- Rule-driven and cost-based optimizer
- Based on Cascades & Large Scope Rules
  - Reduces search space
  - Recognizes patterns such as star joins
- Considers multiple join strategies
  - Nested and nested cache for operational
  - Merge and hybrid hash for large complex queries
- Optimizes inner, outer, & full outer joins
- Considers serial & parallel plans based on cardinality
- Uses equal-height histograms to indicate skew
- Leverages skew buster to eliminate skew
- Un-nests subqueries
- Converts correlated subqueries to joins

- Pushes down predicates to lowest operation
  - Filters e.g. row selection (start-stop key)
  - Coprocessors e.g. pre-aggregation
- Leverages Multi-Dimensional Access (MDAM)
  - To avoid scans when no predicates on leading key columns specified
- Considers sort avoidance strategies
  - Uses hash group by to avoid sorts
  - Leverages key order
  - Does in-memory sort when possible
- Uses sophisticated plan caching techniques
- And a lot more …

*Built & tuned to handle complexities & differences inherent in varied enterprise class workloads*
Why Apache Trafodion™?
Ingredients for a world class RDBMS

3. World Class Parallel Data Flow Execution Engine

- Data Flow pipeline parallel architecture
  - Intermediate results materialized only for blocking operations like sorts
  - Data overflow to disk only for large hash joins
- Adaptive Segmentation to use only needed resources
- Co-located joins & repartitioning when necessary
- Uses Inner and outer child broadcasts
- Parallel secondary index maintenance

Supports salting of data across region servers
Why Apache Trafodion™?
Ingredients for a world class RDBMS

4. World Class Distributed Transaction Management system

- Online backup with point-in-time recovery
- …
Apache Trafodion™ innovation built upon Apache Hadoop™ ecosystem

Leverages Hadoop for core modules

- Hadoop distribution neutral
- Inherited scalability and availability

Differentiation

- Comprehensive ANSI SQL language support
- Relational schema abstraction
- Mature SQL technology with compile and run time workload optimizations
- Automatic query parallelism
- Distributed transaction protection
- Robust data integrity and security enforcement
- Seamless access and integration of Trafodion, native-HBase, and Hive tables
Towards convergence

Stream / Data Flow
- High throughput distributed messaging system
- Enterprise-grade unified stream and batch processing engine
- Distributed realtime computation system
- Fast and general engine for large-scale data processing
- Directed graphs of data routing, transformation, and system mediation logic

App Development
- Framework to build & deploy data applications on Hadoop
- Object/Relational mapping framework

World Class Query Engine
- Apache Trafodion

Development tools
- SQL Client program for any JDBC compliant database
- Universal database tool for developers, DBAs & analysts

In-memory Storage Engine
- In-memory distributed database

Persistent Storage Engine
- distributed, versioned, non-relational database modeled after Google's Bigtable
- smallest, fastest columnar storage for Hadoop workloads
- Distributed graph database
- high-performance, full-featured text search engine library

ETL tools
- Data integration software
- Data Integration & Business Analytics

Development tools
- Helps people see and understand data
- Web-based notebook for interactive data analytics
- Web interface for analyzing data with Hadoop

App Development
- SQL Client program for any JDBC compliant database
- Universal database tool for developers, DBAs & analysts
Performance

YCSB operation speeds that approach Apache HBase™

With max variance at 10.8%
Performance

YCSB and Order Entry scale linearly!

Transaction Order Entry

YCSB

Selects

Updates

19
Performance

Minimum distributed transaction management overhead

**Order Entry**
- multi-statement transactional workload
- 5 transaction types (New Orders, Payments, Order Status, Deliver, and Stock Level checks)
- On average has about 20 statements per transaction

With max variance at 11.3%
Trafodion Open Source

Incubated as open source project by HP Labs and HP IT

Based on 20+ years, $300M investment
Released as open source under the Apache License, Version 2 in June 2014
First “production ready” 1.0 release in January 2015
Follow-on 1.1 release in April 2015
- Includes significant enhancements in performance, manageability, security, high availability, usability
- Throughput at scale reaches Apache HBase and distributed transaction management overhead goals
- More than 2x OLTP improvement with proven linear scalability

Project Trafodion entered Apache Incubator in May 2015

Build an open source community around Apache Trafodion
Apache Trafodion Release 1.3 – January 2016
<table>
<thead>
<tr>
<th>Apache Trafodion</th>
<th>EsgynDB Enterprise</th>
<th>EsgynDB Enterprise Advanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Apache Incubation Project</td>
<td>• Licensed product based on Apache Trafodion</td>
<td>All EsgynDB Enterprise features and versions PLUS</td>
</tr>
<tr>
<td>• Free Open Source</td>
<td>• Hardened, fully tested</td>
<td>• Multi datacenter support</td>
</tr>
<tr>
<td>• Community driven release dates and content</td>
<td>• 24x7 support to meet SLAs</td>
<td>• Future availability, recovery, manageability extensions</td>
</tr>
<tr>
<td>• Community support via email list, voluntary contributions</td>
<td>• Deployment, application design and tuning services</td>
<td>• Future advanced support and services around those features</td>
</tr>
<tr>
<td>• Features as driven by open source community</td>
<td>• Developer, Administrator training</td>
<td>• Additional license charge</td>
</tr>
<tr>
<td></td>
<td>• Enhanced installation and installation support</td>
<td></td>
</tr>
</tbody>
</table>
# EsgynDB Enterprise Capabilities

| Performant                           | • Optimizations for varied workloads  
|                                     | • High level of concurrency          |
| Transactional                       | • Complex                           
|                                     | • Distributed                        |
| Federated                           | • SQL and NoSQL                     
|                                     | • Seamless integration               |
| Secure                              | • LDAP and Kerberos                 
|                                     | • Group and user roles               |
| Extensible                          | • Stored Procedures                 
|                                     | • User-Defined Functions             |
| Flexible                            | • Flexible schemas                  
|                                     | • Multi-structured data              |
| Robust                              | • Data integrity                    
|                                     | • Long-time enterprise usage         |
| Usability                           | • ANSI SQL-99                       
|                                     | • JDBC and ODBC                      |
| Scalability                         | • Start small, grow large           
|                                     | • Linear with added resources        |
| Availability                        | • Failure resilient                 
|                                     | • Online maintenance                |
| Manageability                       | • Query management                  
|                                     | • Resource management               |
| Supportability                      | • Instrumentation                   
|                                     | • Fortune-100 experience             |
Contribute to Apache Trafodion™

Community-led software development

Become a contributor – add a new feature, fix a bug, translate documentation, more

Discuss your changes on the dev mailing list
Create a JIRA issue
Setup your development environment
Prepare a patch containing your changes
Submit the patch
