Anzo Smart Data Lake®

& AnzoGraph

Arthur Keen, Managing Director
Financial Services

http://www.anzograph.com
About

History
2007 – Spun out of IBM’s Advanced Internet Technology Group
2015 – Acquired MPP database Netezza & Paraccel

Leading customers:

Strategic partnerships:

Industry recognition

©2018 Cambridge Semantics Inc. All rights reserved.
Anzo Smart Data Lake®

The industry leading platform for building a

Semantic Layer for the Enterprise

End-To-End  Open Standards  Enterprise Scale
An Open Platform for Big Data Management & Exploration Solutions

- **Automated Ingestion & Mapping**
  - ETL Generation
  - Collaborative Mapping
  - Text Processing

- **Data Cataloging**
  - Data & Model Governance
  - Active Metadata Management
  - Role-Based Security

- **Enterprise Knowledge Graphs**
  - Break Through In-Memory MPP
  - ELT, Model Based Data Prep & Integration
  - Document Search

- **Discovery & Analytics**
  - Automated Query Generation
  - User Dashboard and Custom UI/UX
  - Self-Serve Live Extracts

**CONNECTORS** → **MODELS** → **RULES** → **ANALYTICS & TOOLS**
The Three Rings:
BlazeGraph: Meta-data management
Spark: Transformation
AnzoGraph: Graph OLAP at Scale
Shared Distributed Storage
HDFS, S3, GCS, Azure Storage,...

BlazeGraph
(GOLTP)

Apache Spark
(Transform Analyze)

AnzoGraph
(GOLAP)
Middle Ware:
Anzo Seamlessly integrates BlazeGraph, Spark, AnzoGraph

BlazeGraph (GOLTP)

Apache Spark (Transform Analyze)

AnzoGraph (GOLAP)

HDFS/S3/...
**AnzoServer Roles**

**Manage** Metadata, Provenance, Dashboards, Users, Access Control, Configuration, Execution, Services, Directory, ...

**Generate** Sparql, Spark Scala, and Models.

---

**BlazeGraph** (GOLTP)

**AnzoServer**

**HDFS/S3/...**

**Apache Spark** (Transform Analyze)

**AnzoGraph** (GOLAP)

---

**SPARQL http/GRPC**

**Consume**

**Consume Prepared Data**

Anzo Dashboards
3rd Party BI (OData, WDC)
Databases
Custom Apps

---

**Rest**

**OData**
**AnzoServer Roles**

- **Manage** Metadata, Provenance, Dashboards, Users, Access Control, Configuration, Execution, Services, Directory, ...
- **Generate** Sparql, Spark Scala, and Models.

**Spark Roles**
- Transformation
- Analytics
- Graphs

**Connections**

- **Apache Spark** (Transform Analyze)
- **BlazeGraph** (GOLTP)
- **AnzoGraph** (GOLAP)

**AnzoServer**

- Manage Metadata, Provenance, Dashboards, Users, Access Control, Configuration, Execution, Services, Directory, ...
- Generate Sparql, Spark Scala, and Models.

**Consume Prepared Data**
- Anzo Dashboards
- 3rd Party BI (OData, WDC)
- Databases
- Custom Apps

**SPARQL http/GRPC**
- Rest
- OData

**Consume**
- Prepared Data

**HDFS/S3/...**

**Graphs**
AnzoServer Roles

**Manage** Metadata, Provenance, Dashboards, Users, Access Control, Configuration, Execution, Services, Directory, ...

**Generate** Sparql, Spark Scala, and Models.

Spark Roles
- Transformation
- Analytics
- Graphs

AnzoGraph Roles
- Complex Query Performance at Scale (Load, Query, Update, Persist, UDF)

AnzoGraph (GOLAP)

Apache Spark (Transform Analyze)

BlazeGraph (GOLTP)

HDFS/S3/...

Consume Prepared Data
- Anzo Dashboards
- 3rd Party BI (OData, WDC)
- Databases
- Custom Apps

Consume

Connections

SPARQL http/GRPC
- Rest
- OData

Generate Sparql, Spark Scala, and Models.
AnzoGraph

- Massively Parallel Processing (MPP)
  - Designed for Analytics (join vs. traverse)
  - Shared Nothing
  - Each core contributes to each query
- Ancestry Paracel, Redshift, Neteeza
  - Same lead developers from Paracel
  - Perpetual license to some of structural code
  - Re-implemented for Graph
- Benchmarks
  - 2016 LUBM Trillion-Triple
  - 2018 TPC-H 1000
AnzoGraph

- Massively Parallel Processing (MPP)
  - Designed for Analytics (join vs. traverse)
  - Shared Nothing
  - Each core contributes to each query
- Ancestry Paraccel, Redshift, Neteeza
  - Same lead developers from Paraccel
  - Perpetual license to some of structural code

Late-breaking news: Cray has posted Trillion Triple numbers June 2018.

<table>
<thead>
<tr>
<th>Graph Database</th>
<th>Load Time</th>
<th>Inference Time</th>
<th>Query Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oracle Database 12c [2014]</td>
<td>115.2 hours</td>
<td>86.5 hours</td>
<td>22.5 hours</td>
</tr>
<tr>
<td>Cambridge Semantics ANZO Graph Query Engine [2016]</td>
<td>1,760 seconds</td>
<td>4,574 seconds</td>
<td>840 seconds</td>
</tr>
<tr>
<td>Cray Graph Engine [2018]</td>
<td>4,124 seconds</td>
<td>535 seconds</td>
<td>96 seconds</td>
</tr>
</tbody>
</table>
AnzoGraph Strengths

• Very High Speed Load
  • 250 GB/hour/node (CSV on 32vCPU nodes)
  • Fully parallel
  • Cores over-subscribed in pipeline
• Complex Queries
  • Cost/Rules hybrid query planner
• Extended SPARQL/RDF
  • Window functions
  • 70+ added “Excel” builtins, including aggregates
• Views
• Auto-index & compression
  • Background Vacuum & Index optimization
Highlights

- Generated code (C++)
  - Runs “close to the silicon”
  - Eliminates interpreter overhead
  - Re-uses previous code snippets
  - Educated by Netezza, Paraccel
- Long pipelined flows
  - Non-materializing across (multiple) network hops
  - Each query uses multiple threads on every core
- Network Model
  - Throughput, not latency (dataset “push”)
- Dictionary-based compression
  - Late decompression
Current Projects

• Peta-Scale Scanning
  • HDFS, Local drives, S3
• RDF* “RDR - Reification Done Right”
• Property Graphs
• Tinkerpop Gremlin Support
Questions?

Try out AnzoGraph:
http://www.anzograph.com