CAMBRIDGE SEMANTICS
THE SMART DATA COMPANY®

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:

Lilly, MERCK, Boehringer Ingelheim, NOVARTIS, IEEE, BMO, CREDIT SUISSE, BRIDGEWATER, SANOFI, Bristol-Myers Squibb, Biogen, pwc, KPMG, STAPLES, Broadridge, Vanguard, Jefferies, WELLS FARGO

Strategic partnerships:

pwc, Google Cloud Platform, THOMSON REUTERS, INOVA, Linguamatics, SciBite The language of science, citeline an informa business

Industry recognition

//CODiE// 2017 SIIA CODiE WINNER

DBTA 100

2017 STEVIE SILVER WINNER AMERICAN BUSINESS AWARDS

KM World 100 COMPANIES That Matter in Knowledge Management

KM World Trend-Setting PRODUCT OF 2017

CIO 100 MOST PROMISING BIG DATA Review SOLUTION PROVIDERS·2017

Anzo Smart Data Lake®

The industry leading platform for building a

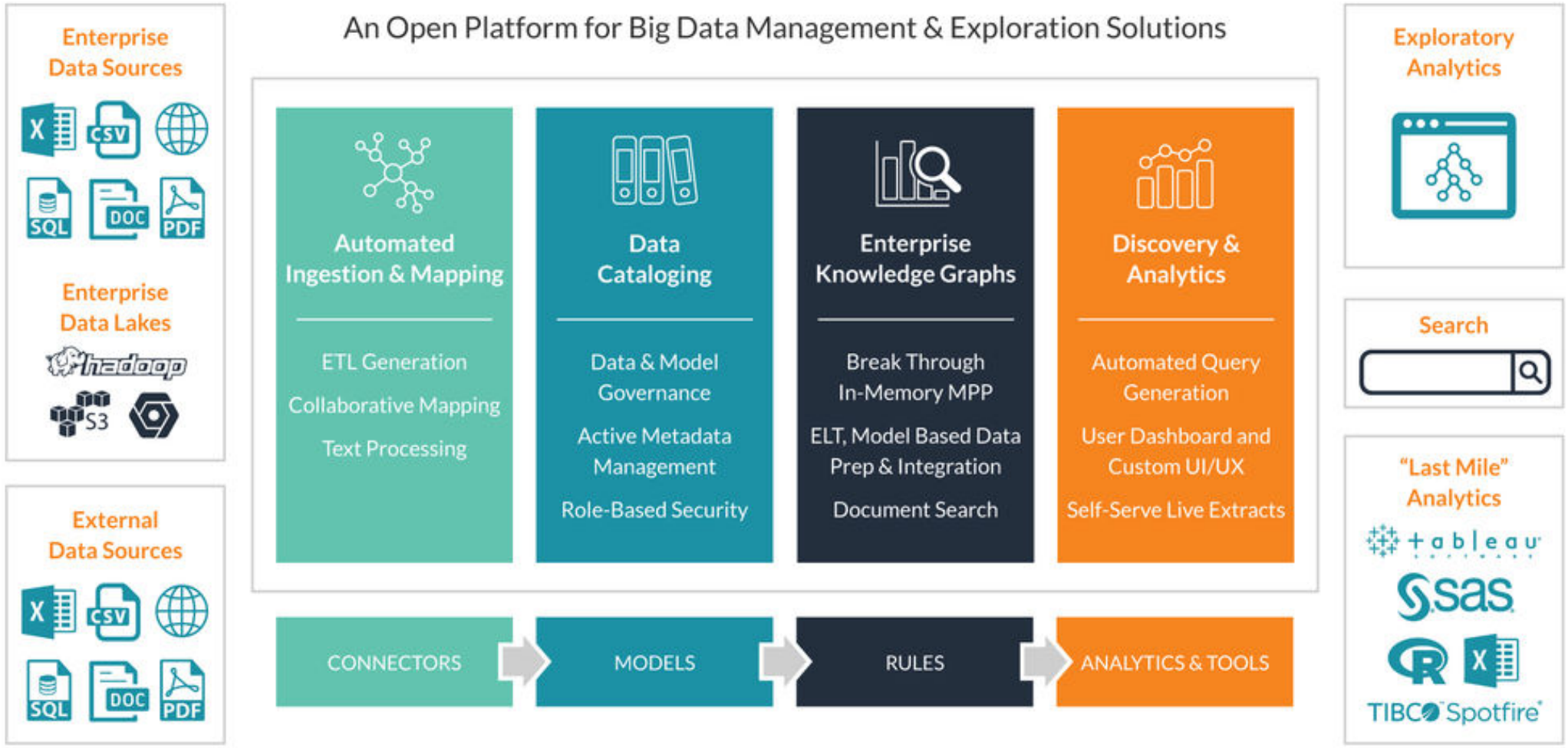
Semantic Layer for the Enterprise

End-To-End

Open Standards

Enterprise Scale

Anzo Smart Data Lake 4.0



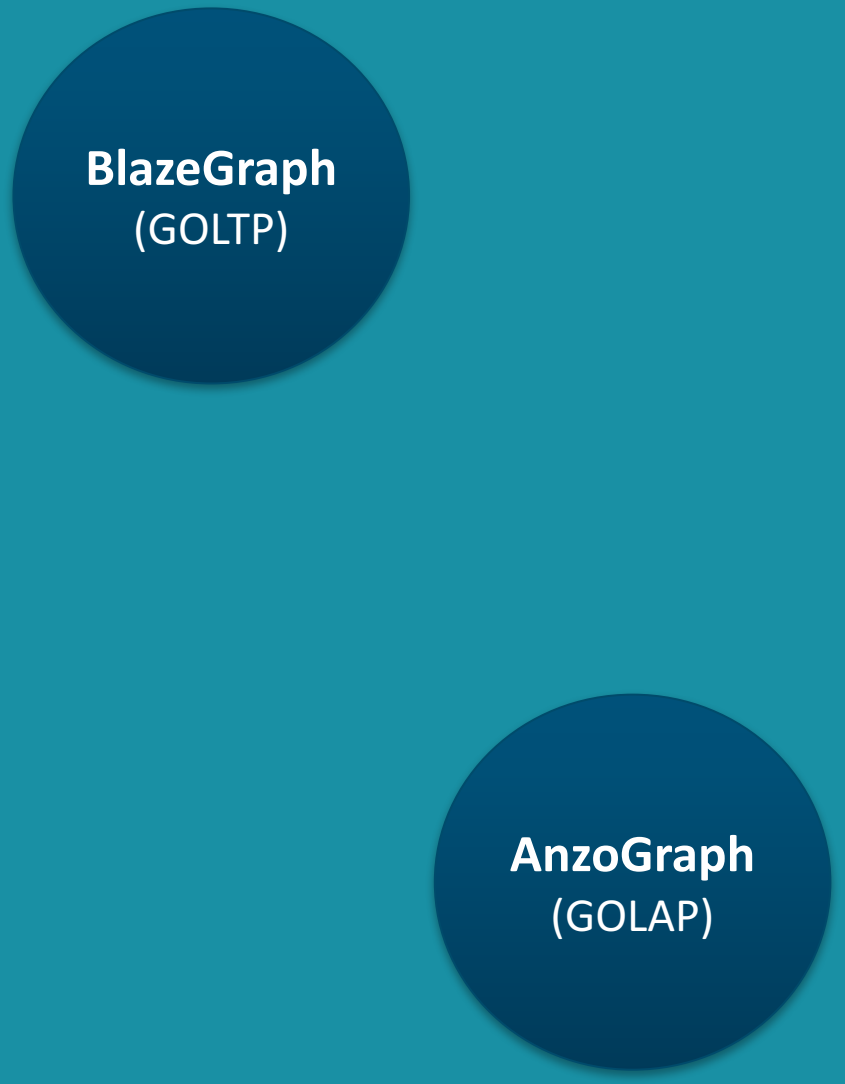

THREE RING CIRCUS
MASTER CLOWNS SHOW
CIRCUS
BALL
1.
2.
3.
Richard Capes

The Three Rings:

BlazeGraph: Meta-data management

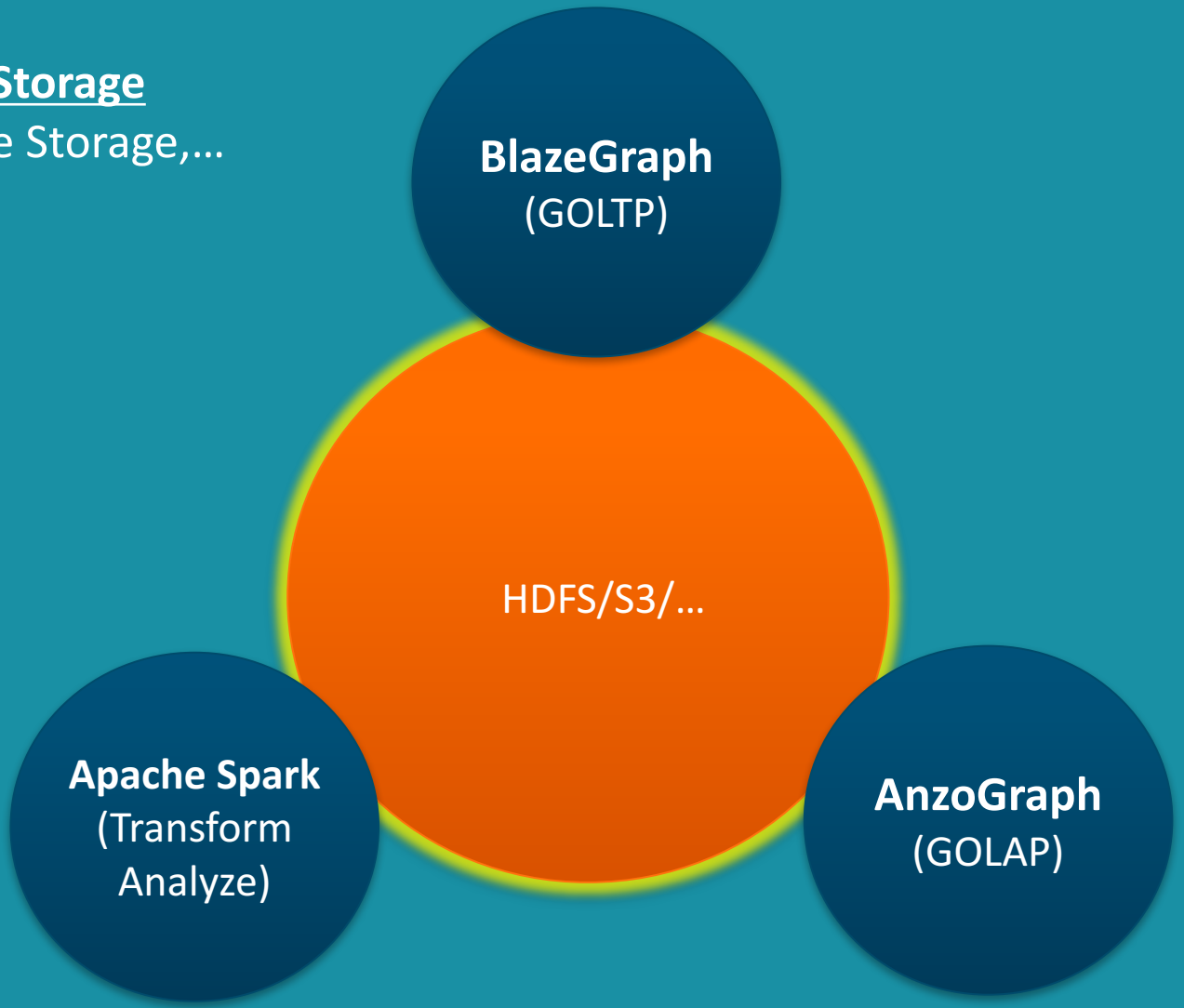
Spark: Transformation

AnzoGraph: Graph OLAP at Scale



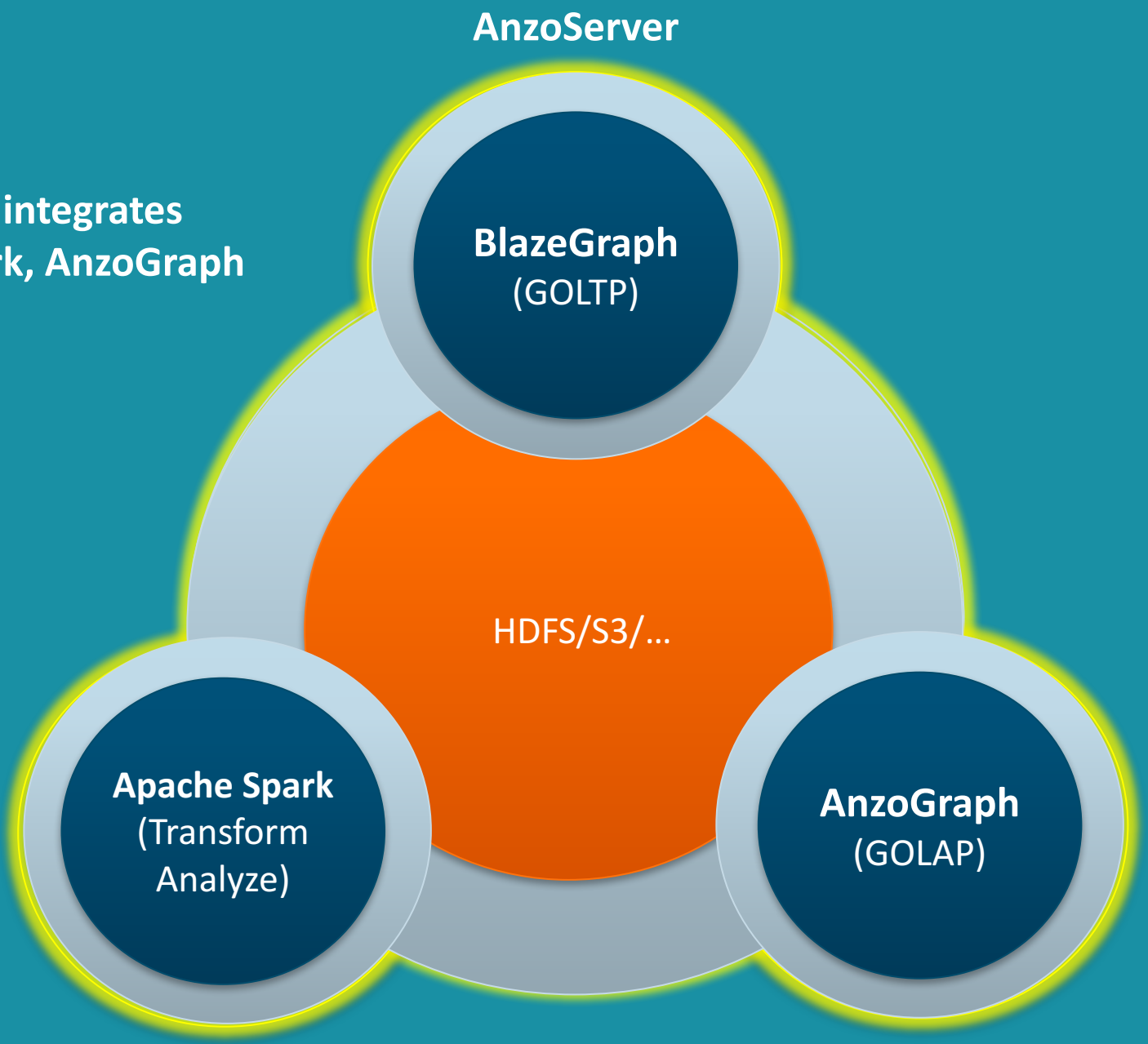
Shared Distributed Storage

HDFS, S3, GCS, Azure Storage,…



Middle Ware:

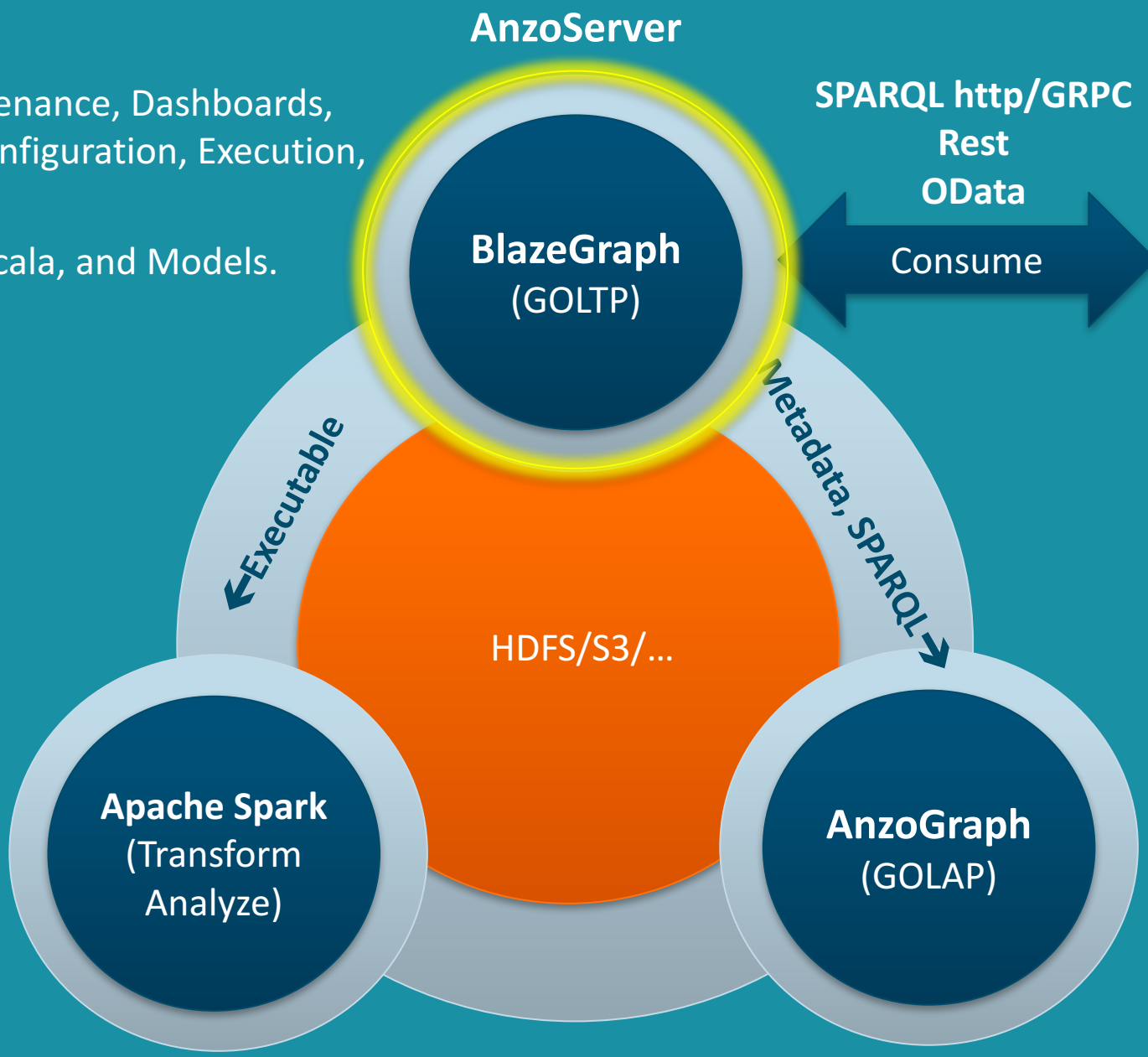
Anzo Seamlessly integrates BlazeGraph, Spark, AnzoGraph



AnzoServer Roles

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

Generate Sparql, Spark Scala, and Models.



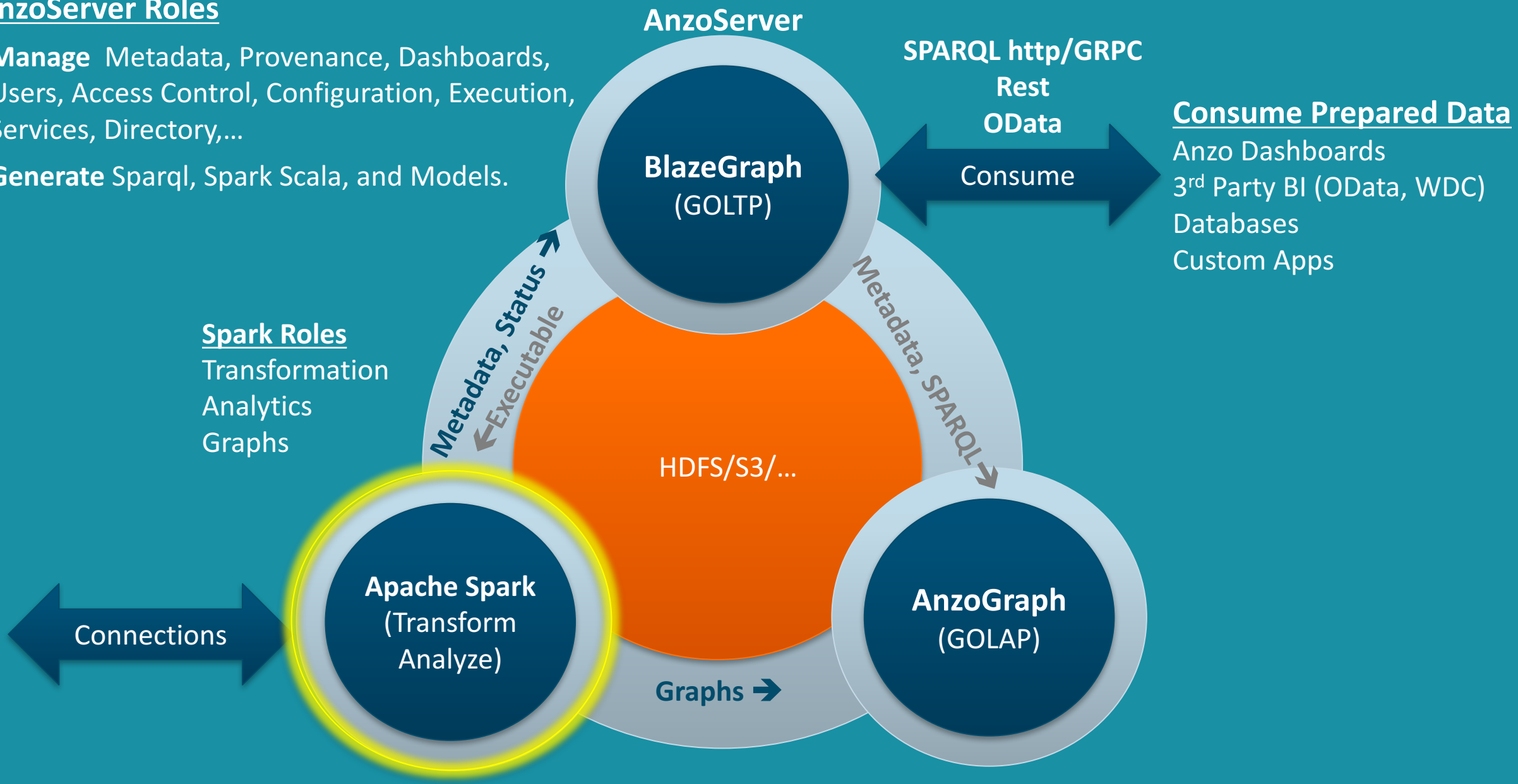
Consume Prepared Data

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

AnzoServer Roles

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

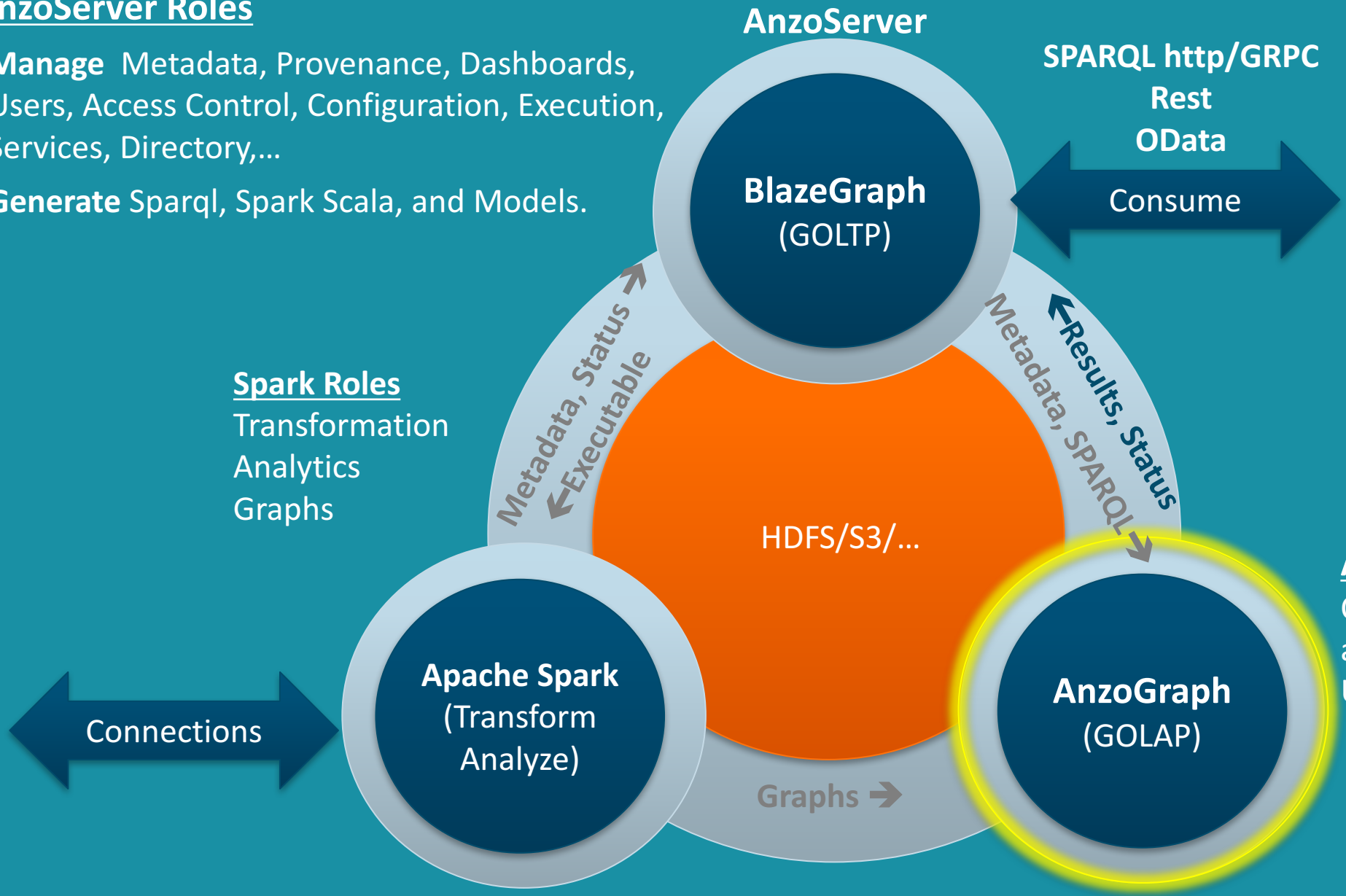
Spark Roles

Transformation
Analytics
Graphs

AnzoServer Roles

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

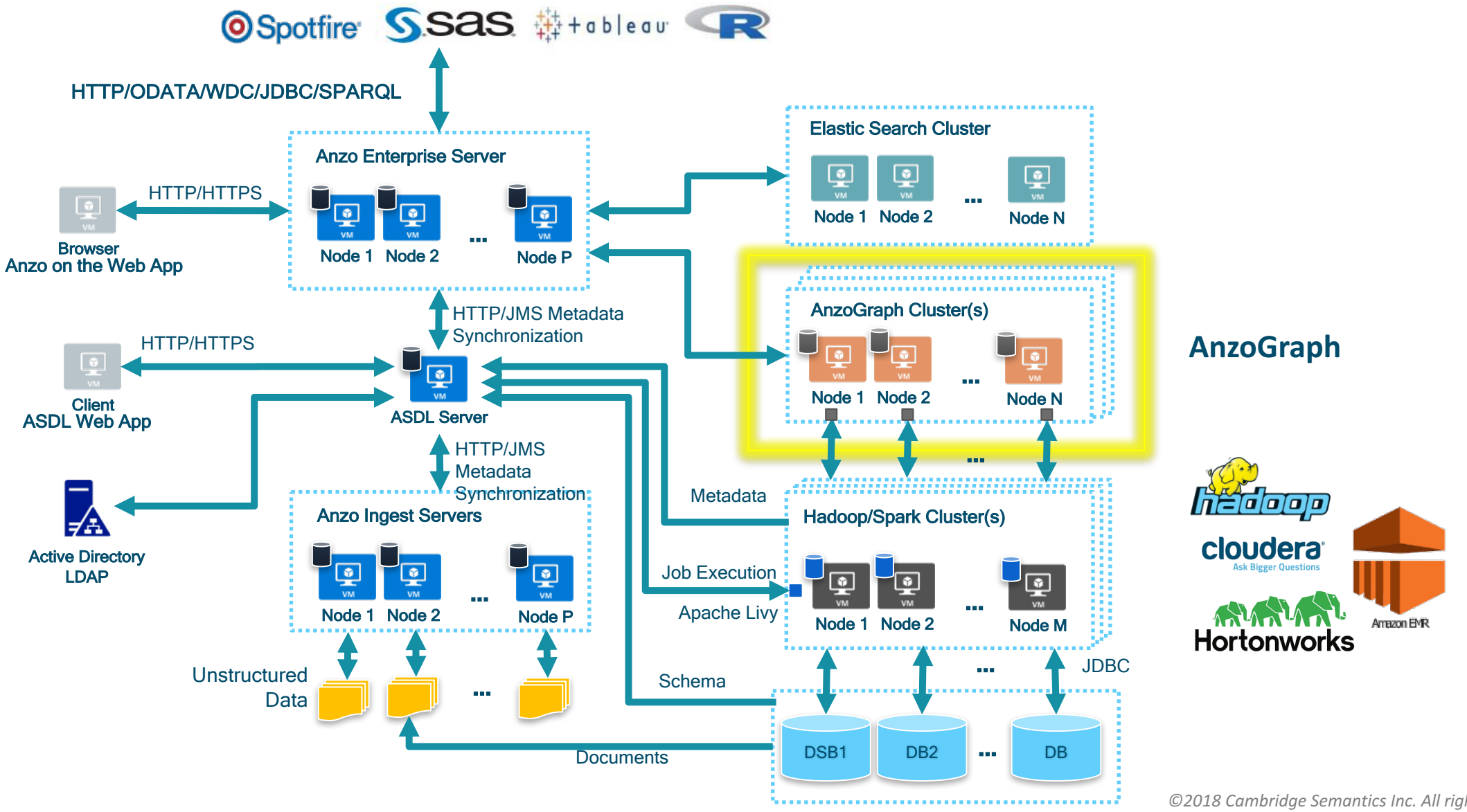
Spark Roles

Transformation
Analytics
Graphs

AnzoGraph Roles

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

Reference Architecture



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
  - 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.

Graph Database	Load Time	Inference Time	Query Time
Oracle Database 12c [2014]	115.2 hours	86.5 hours	22.5 hours
Cambridge Semantics ANZO Graph Query Engine [2016]	1,760 seconds	4,574 seconds	840 seconds
Cray Graph Engine [2018]	4,124 seconds	535 seconds	96 seconds

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