



BIG DATA APPLIANCES

July 23, TDWI

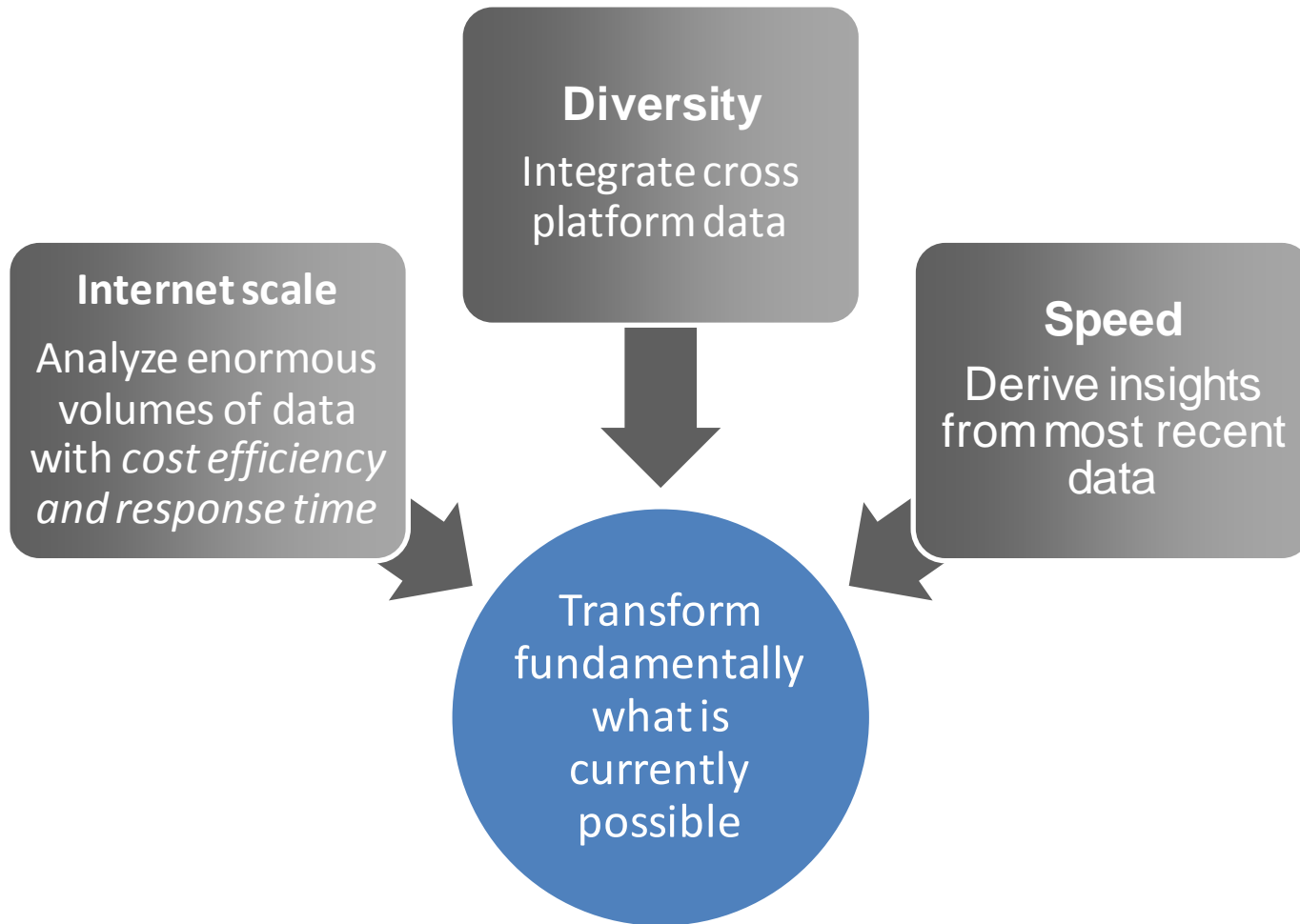
R Sathyanarayana

Enterprise Information Management &
Analytics Practice

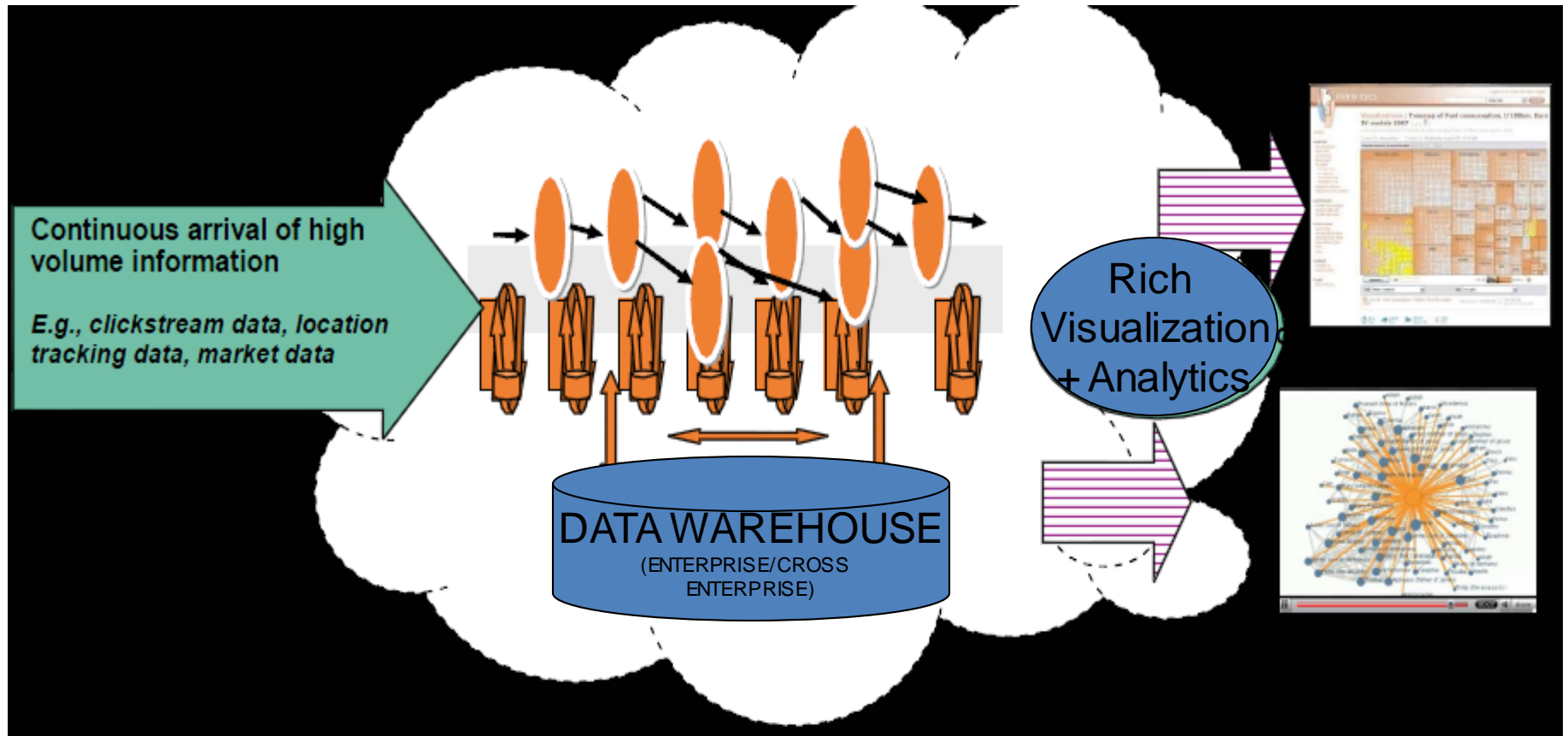
EMC Consulting

Big data are datasets that grow so large that they become awkward to work with using on-hand database management tools. Difficulties include capture, storage, search, sharing, analytics, and visualizing. This trend continues because of the benefits of working with larger and larger datasets allowing analysts to "spot business trends, prevent diseases, combat crime"

What's all the fuss about 'Big Data'?



Picture This....



Appliances : Setting the context

As IT organizations build up massive numbers of databases to deal with the explosion of data, the ability to make real-time decisions on new questions (BI) that involve enormous amounts of information (DW) will need to be a core competency for many organizations.

Due to this shift, DW/BI customers need a solution that can provide extreme predictable performance, scale-out architecture for 'Big Data' analytics and an enterprise-proven feature set all at the lowest TCO.

Under Works :

Definition of Datawarehouse Appliance

Original Definition : Hardware + Software – built and supported by a single vendor

Partial Technology Stack : May or may not bundle with other vendors' hardware and / or operating system

Architectural considerations are evolving

	Whole Technology Stack (Appliances or Bundles)	Partial Technology Stack	Miscellaneous
Data Warehouse Appliances (DWAs)	<p>Early DWAs:</p> <ul style="list-style-type: none"> • Netezza (now IBM) • Teradata <p>New DWAs:</p> <ul style="list-style-type: none"> • EMC Greenplum DCA • XtremeData dbx <p>Very Early Products:</p> <ul style="list-style-type: none"> • Sequent, White Cross 	<p>Relational Databases:</p> <ul style="list-style-type: none"> • Aster MapReduce DWA (now Teradata) • Greenplum (EMC DCA) • Kickfire (now Teradata) • Kognitio (SaaS) <p>Columnar Databases:</p> <ul style="list-style-type: none"> • ParAccel, Sybase IQ, Vertica (now HP) 	<p>General Data Appliance:</p> <ul style="list-style-type: none"> • Dataupia Sartori Server • Oracle Exadata (storage) <p>Add-On Accelerators:</p> <ul style="list-style-type: none"> • Cognos Now! • SAP BI Accelerator • Teradata Accelerate
Related Products	<p>Server/Database Bundles:</p> <ul style="list-style-type: none"> • IBM Balanced Warehouse • Oracle Database Machine • SQL server Fast Track • Sybase Analytic Appliance 	<p>Open-Source Databases:</p> <ul style="list-style-type: none"> • Infobright • Ingres • Postgres • MySQL <p>Open-Source Op Sys:</p> <ul style="list-style-type: none"> • LINUX (various vendors) 	<p>Commodity/Standardized Hardware:</p> <ul style="list-style-type: none"> • Dell, HP, IBM • Cisco, EMC • AMD, Intel

Trends: Consolidation in the industry

Focus is shifting in multiple areas:

1. From whole technology stack to pieces of it
2. From hardware to software
3. From proprietary to commodity hardware
4. From new vendors to infrastructure providers
5. From single to mixed workloads
6. From data marts to EDW

Major differences between the DW appliances

- Column vs. Row Storage
- Polymorphic Storage (Both Column and Row)
- Proprietary and Commodity Hardware
- In-Memory Processing
- Relationship with Existing Architecture
- Shared Nothing Architecture

The Challenges in Today's Data Warehousing Environments

- Sources of data and the amount of data to analyze is growing exponentially
- Stale data exists because DW solutions cannot ingest the vast amounts of data fast enough
- Lack of performance for advanced analytics and complex queries
- The number of users and the concurrency of users is increasing rapidly

Top questions from architectural viewpoints

- Proprietary / Non-proprietary, standardized hw
- Is this a shared nothing / share disk/ share everything architecture?
- Capability – in database analytics?
- Scale – linear scalability?
- What are your ingestion rates? (higher rate, better real time reporting)
- Can I balance my workloads? How about mixed workloads?
- What are my DW management overheads?

Considerations of DW Solution

- Easily scales to analyze growing amounts of data
- Rapidly ingests large amounts of data from sources
- Provides high performance in database analytics
- Supports high user concurrency securely, reliably
- Handle multiple workloads