Welcome to the Virtual IMS user group newsletter. The Virtual IMS user group at www.fundi.com/virtualims is an independently-operated vendor-neutral site run by and for the IMS user community.

Virtual IMS user group presentation

The latest webinar from the Virtual IMS user group was entitled, "Best Practices – Populating Big Data Repositories from IMS", and was presented by SQData's CEO, Scott Quillicy.

Scott is the CEO and founder of SQData, an Addison, TX-based software company that specializes in high-performance data movement and changed data capture (CDC) for IMS and relational databases. He has over 30 years of database experience and is considered an expert in database replication strategy and deployment. Scott has managed a number of large data integration projects, many of which involved the near-real-time data capture of IMS databases to a variety
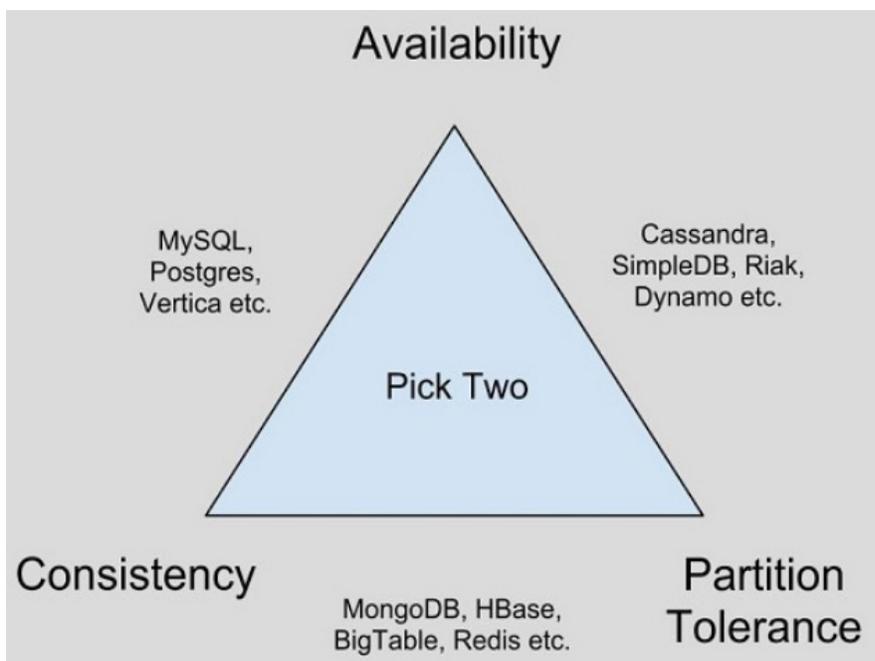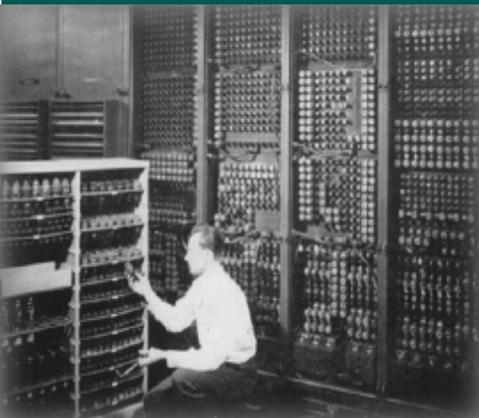

Figure 1: CAP theorem

of downstream databases / applications.

Scott started the presentation by exploding certain Big Data myths. It's not the 'New Wave' of technology, nor is it exclusively Hadoop and/or NoSQL-based. It's
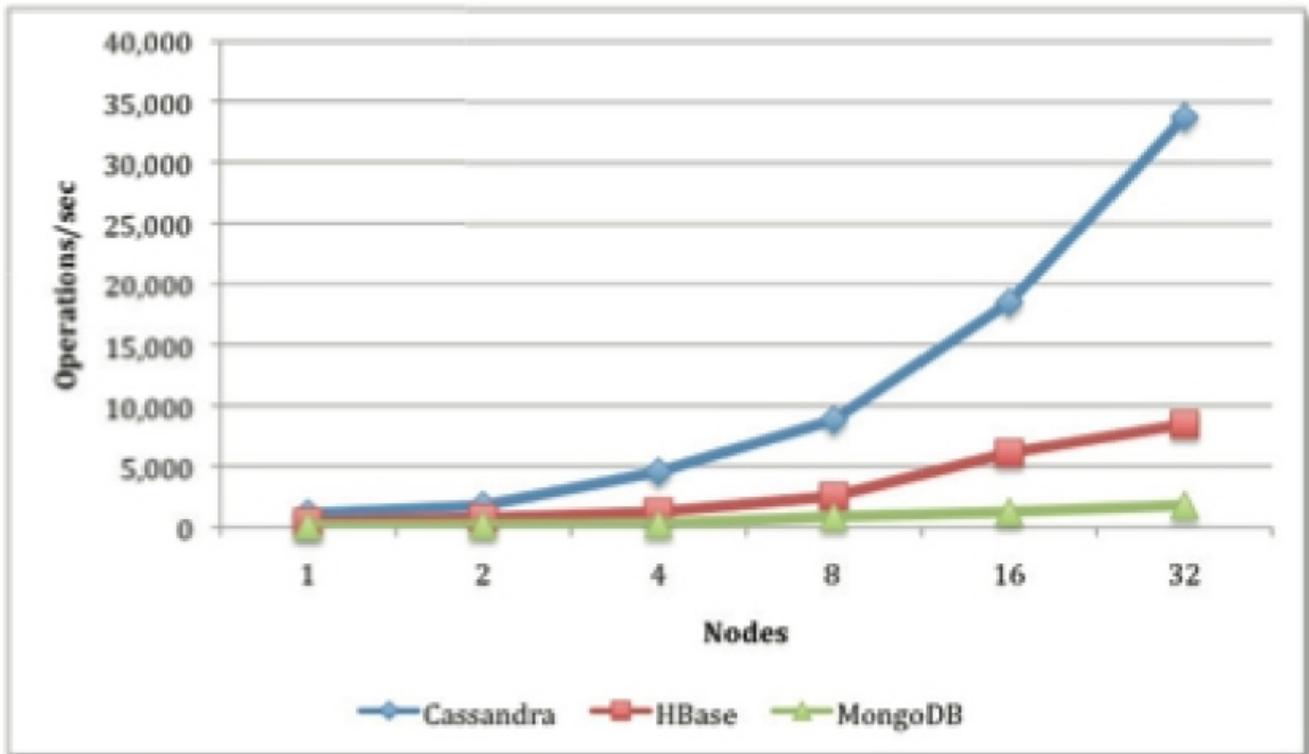
Contents:

Figure 2: Performance – read/write mix workload

not advanced analytics of disparate data, nor does Big Data 'know' what you are doing. What it is, is a large collection of data, and has been around for over 50 years.

The characteristics of Big Data are that it involves significant amounts of data, in many different formats, with high rates of change, and it's complex.

It comes with certain challenges, such as the increasing data volumes will stress traditional RDBMSs, and there are computing and infrastructure costs to

process / analyse the data

Scott informed the user group that there were more choices, when it came to Big Data, than Hadoop and NoSQL.

He first introduced the Hadoop family, saying that HDFS (Hadoop Distributed File System) is the basic file system. HBase is a NoSQL database built on HDFS. HCatalog contains the metadata. Hive provides an SQL interface. And Pig is a scripting language used for MapReduce for unstructured sources.

The first of the main alternatives he mentioned is Cassandra. This is a wide-column store that handles very large datasets in "almost" SQL. It has a ring architecture, providing selectable replication. The other main alternative is MongoDB. It's a popular document store using JSON / BSON format for master / slave replication.

Scott Quillicy reminded everyone of Eric Brewer's 1998 CAP Theorem (see Figure 1). This states that it's impossible for a distributed system to provide all three
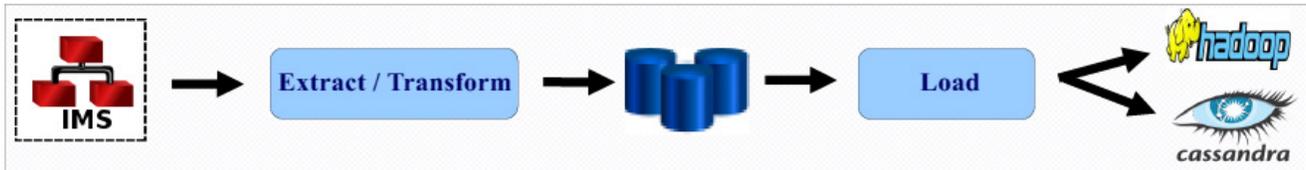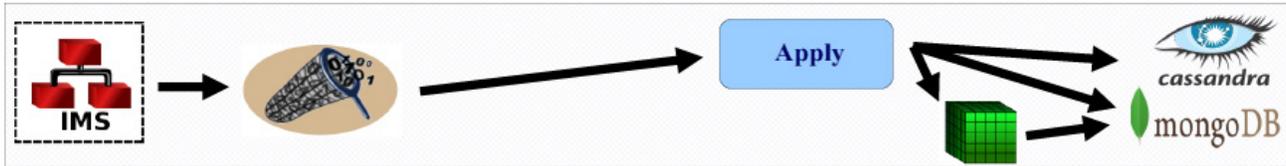
Figure 3: Extract, Transform, Load

Figure 4: Change Data Capture

guarantees of availability, consistency, and partition tolerance.

Choosing which database to use depends on which two of the three points on the triangle are most important for the kind of work that you'll be doing. The other consideration is how well the databases scale – see Figure 2.

Why would you move data from IMS to Big Data? Scott suggested that it provides a way to analyse data outside of IMS using Business Intelligence / advanced analytics. The data can be combined with data from other applications, both structured and unstructured. The data can also be moved to inexpensive computing / storage. And it compliments established Data

Warehouse(s). And Scott added that the good news was that the process is less complicated than moving data from IMS to relational databases.

Scott provided some 'best practice' hints and tips. He said to let the business drive the effort by ensuring proper alignment with business goals, to let queries drive the data model design, avoiding the IT-initiated 'build it and they will come' approach

He suggested that sites should temper their exuberance, warning that this is inevitable after the successful implementation of an application. He also warned that technology is rapidly evolving, so what is OK today may be obsolete tomorrow. Adding that it is more expensive than the

hype leads you to believe.

Scott advised that the project should align with enterprise data, and this is where IT takes a lead role because they know about existing Data Warehouse / Business Intelligence set-ups, and the existing infrastructure should fit with the data integration.

Scott recommended using an iterative approach for implementation, such as Agile or Agile-like, and to set aside the relational mindset . And to allow for 'adjustments' without them having a major impact on the schedule.

The key considerations are:

• Big Data repository selection – consider Open Source projects with large communities; beware of vendor lock-in; and more

than one may be required for the project.

- Data delivery / latency – should be business driven; do you need full extracts or periodic extracts; is it near-Real-Time or scheduled changes.

- Workload characteristics – read versus update ratio; consider the update volume and changes as a percentage of a particular source. These will affect the choice of Big Data repository.

- Format – level of normalization (less is usually desirable); privacy / masking; and the level of transformation.

The usual challenges exist when transforming data from one format to another. When it comes to ETL (Extract, Transform, Load), the considerations are: whether you need full data extract / load; is data transformation logic defined in this step; is it an iterative process (which must be fast and efficient}; and it should minimize the amount of data landing. See Figure 3.

The considerations for CDC (Changed Data Capture) include: move only the data that has changed; it's ideal for sequence of events; re-use data transformation logic

from ETL; use near-real-time / deferred latency.

A copy of Scott's presentation is available for download from the Virtual IMS user group Web site at www.fundi.com/virtualims/presentations/IMStoBigDataOct14.pdf.

You can see and hear the whole user group meeting by downloading the WMV file from www.fundi.com/virtualims/presentations/2014-10-07meeting.wmv.

Meeting dates

- On 9 December 2014, we have Rocket Software's Wayne Morton.

- On 10 February 2015, it's a speaker from DataVantage.

Recent IMS articles

Replicate Data Easily With InfoSphere Classic Change Data Capture for z/OS by Karen Durward in *IBM Systems Magazine (August 2014).* You can find the article at `http://www.ibmsystemsmag.com/mainframe/administrator/ims/Classic-CDC/`.

IMS news

Syncsort has announced new technology, which is now committed as SQOOP-1272, that supports loading multiple mainframe data sets to each of the nodes in a Hadoop cluster in parallel and transforming them into any Apache Sqoop-supported file format. This makes it simple for organizations to integrate data from mainframe databases, such as DB2/z, IMS, Adabas, IDMS, and Datacom, with the rest of the data in a typical next-generation Big Data environment. This make it more cost-effective to store mainframe historical data in HDFS and will also help free-up mainframe CPU cycles by allowing customers to move expensive data processing workloads from the mainframe to Hadoop. More information can be found at `www.syncsort.com/en/About/News-Center/Press-Release/Syncsort-Contribution-to-Apache-Sqoop`.

GT Software has announced Version 4.6 of Ivory Service Architect, which allows users to extend access to mainframe applications, data, and processes through Web services. It now provides support for IMS/TM multiple segment output messages, and variable

length output messages. More information can be found at www.gtsoftware.com/gt-software-announces-new-version-ivory-service-architect-supporting-enhanced-enterprise-mainframe-scalability-performance-security/.

BMC Software has announced BMC Subsystem Optimizer for zEnterprise, which removes the traditional requirement that DB2, IMS, and CICS subsystems must reside on the same LPAR to communicate and process workloads. Users can now reduce MLC charges by reconfiguring their subsystem placement across partitions. The product ensures that calls to DB2 and IMS can always be delivered to a subsystem on other partitions. More information can be found at www.bmc.com/news/press-releases/2014/BMC-Subsystem-Optimizer-Reduces-Mainframe-Software-License-Costs-.html.
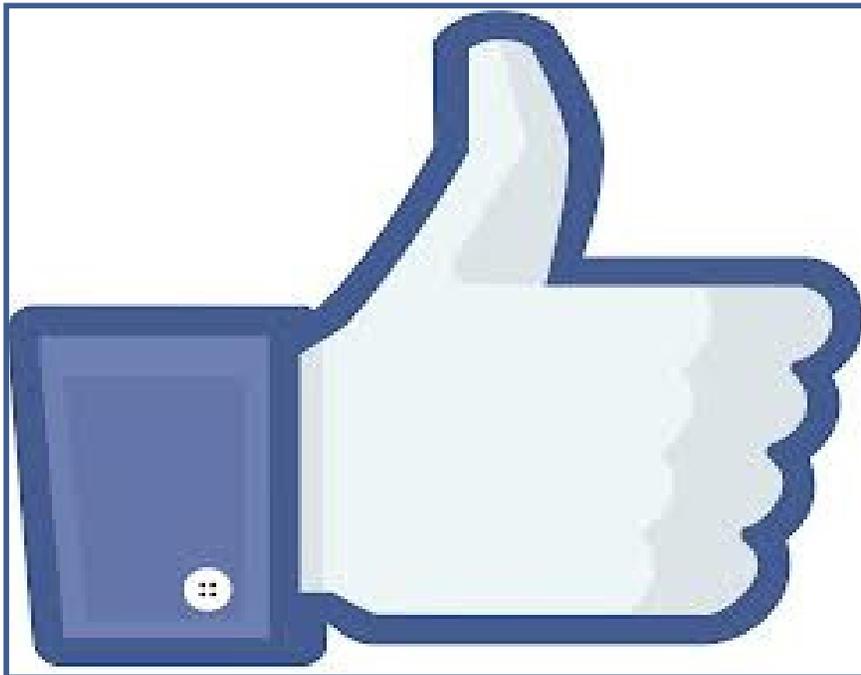
Like us on Facebook

About the Virtual IMS user group

The Virtual IMS user group was established as a way for individuals using IBM's IMS hierarchical database and transaction processing systems to exchange information, learn new techniques, and advance their skills with the product.

The Web site at www.fundi.com/virtualims provides a central point for coordinating periodic meetings (which contain technically-oriented topics presented in a webinar format), and provides articles, discussions, links, and other resources of interest to IBM IMS practitioners. Anyone with an interest in IMS is welcome to join the Virtual IMS user group and share in the knowledge exchange.

To share ideas, and for further information, contact trevor@itech-ed.com.

The Virtual IMS user group is free to its members.