# Virtual IMS user group: Newsletter 57

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, “Real-Time Streaming – IMS to Apache Kafka and Hadoop”. It was presented by SQData’s Scott Quillicy.

Scott is a co-founder of SQData, a US-based based software company providing high-performance data replication and changed data capture (CDC) for IMS, VSAM, DB2, and distributed relational databases such as Oracle and DB2 LUW. With over 35 years of database experience, Scott is considered an expert in database replication strategy and deployment. Lately, Scott spends much of his time with real-time streaming of mainframe data into Big Data platforms.

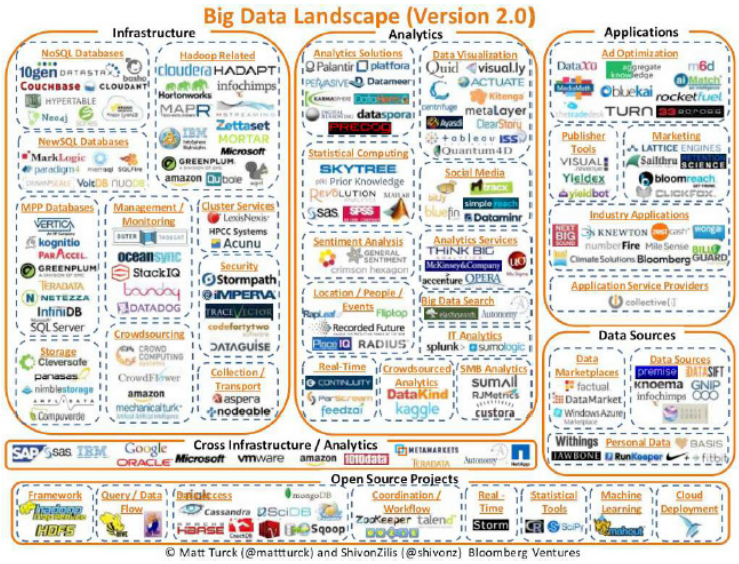


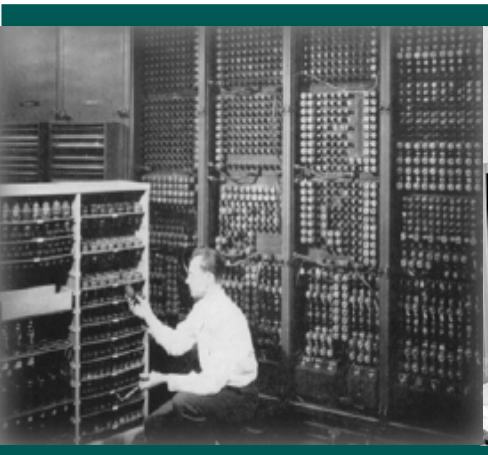

**Figure 1: Big Data choices**

Scott Quillicy started his presentation by suggesting that when it comes to Big Data, you have plenty of choice (with more on the way) – see Figure 1.

Although Big Data is a new development, the reality is that a large collection of data has been in existence

***Contents:***





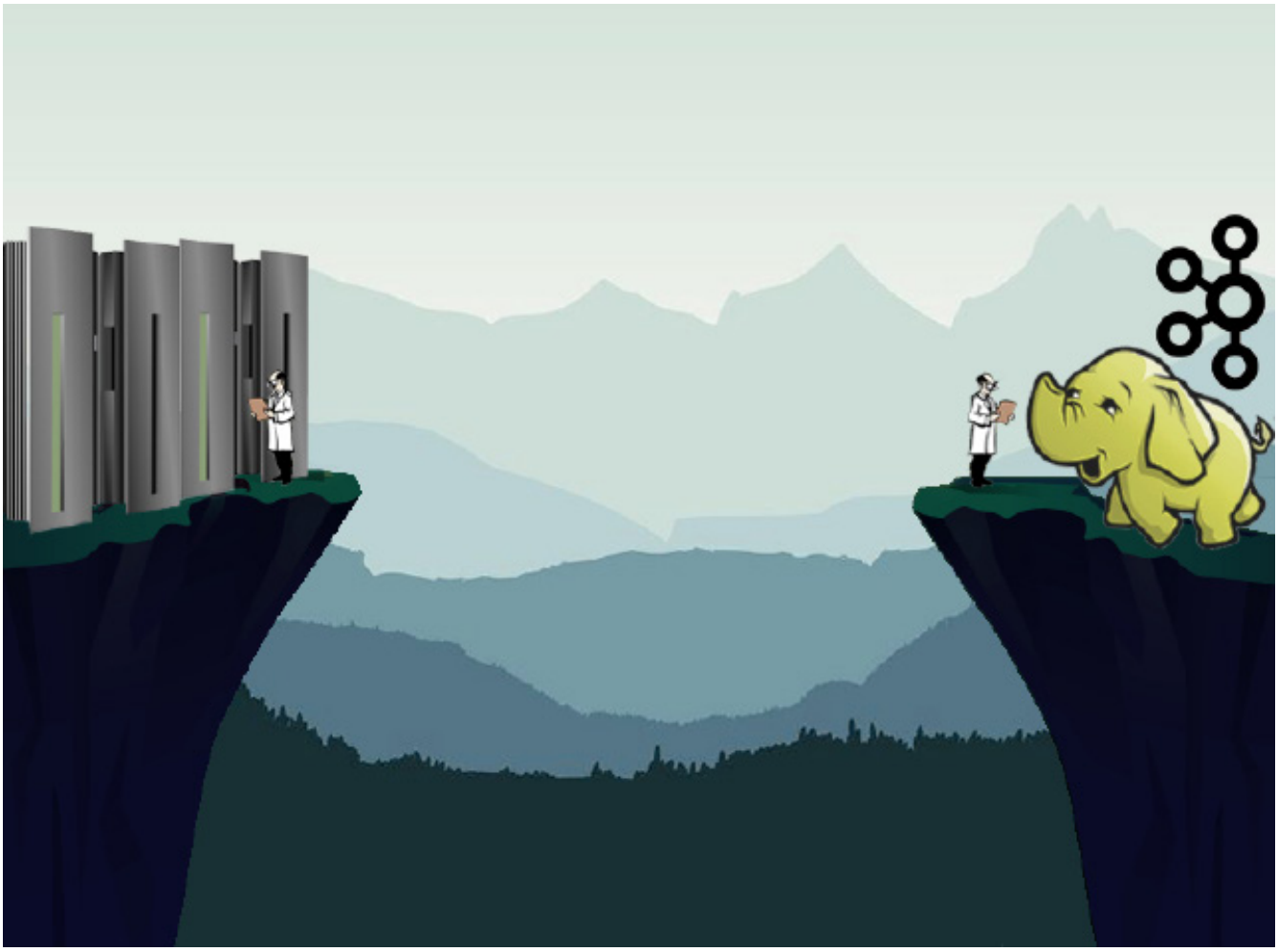


**Figure 2: The great divide**

on mainframes for over 50 years. The characteristics of Big Data are that there is a significant amount of data that allow advanced analytics of disparate data, which comes in many different formats (structured, semi-structured, un-structured), and there's a high rate of change.

The challenges that sites are facing include increasing data volumes, which can stress traditional RDBMS, and the computing and infrastructure costs to process / analyze the data. There are exciting times ahead because there is a large open source community and there's rapid evolution of the technology.

Why are sites looking at real-time streaming of mainframe data to Big Data? The answer is analytics. It's the difference between making decisions based on current information against data that's 24 hours or more out of date. That means users can quickly detect key events / trends, and it helps the organization maintain a competitive advantage and provide better customer service, which in turn increases revenue and profitability.

Scott went on to compare real-time against ETL (Extract Transform Load). He said that an IDC study found that nearly 2/3rds of the data moved by ETL was at least 5 days old before reaching an analytics database. The survey revealed that it takes at least 10 minutes to move 65% of CDC (Change Data Capture) data into an analytics database. 75% of IT executives worry about

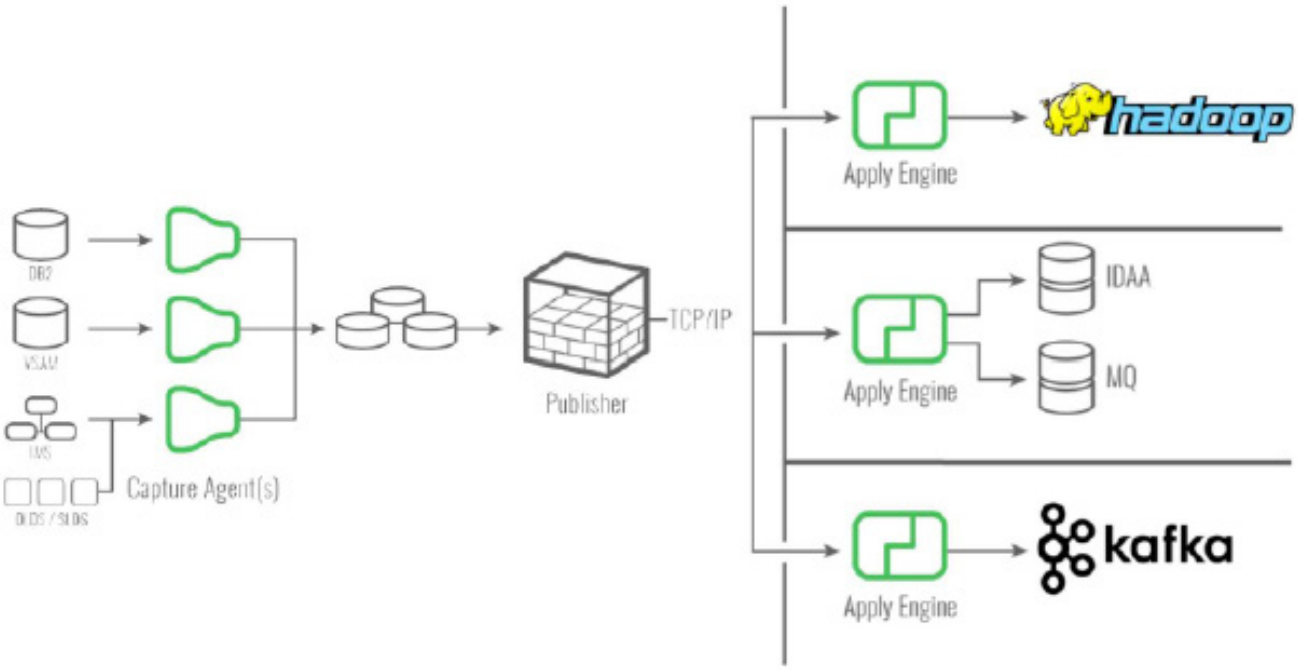



**Figure 3: Mainframe data streaming**

data lag that might hurt their business. And 27% said data disconnect is slowing productivity.

Figure 2 illustrates the great divide that occurs at many sites between the mainframers and the Big Data people.

So, what are the popular big data products? Hadoop HDFS is the most commonly used Big Data store. It can be the foundation for other technologies (eg Spark). And it's highly scalable.

Hbase is a NoSQL key-value store. Tables split into column families. It allows for inserts and updates. And it's intended for real-time queries

Hive is a data warehouse infrastructure built on HDFS. It allows data stored on HDFS to be queried. It runs only in batch. And its main use is for analyzing data collected over time.

Kafka is an ultra-fast message broker. It streams data into most popular Big Data targets. It can have multiple producers / consumers. And it is ideal for real-time streaming.

Other popular stores available include Cassandra, MongoDB, and Spark.

Just migrating to Big Data isn't the answer. Scott highlighted the major mistakes that some sites were making.

Some sites, for example, didn't have a clear use-case(s). Their idea was to build it and they will come – assuming that people would use it just because it was there. Scott saw this as a great way to ensure failure because there was minimal focus on business needs and it often happened because the staff were under pressure to deploy some kind of Big Data solution.

Other sites fail because of data collection overkill. They think that everything should be in data lakes. This approach wastes time moving data that has little business value. It guarantees timeline and cost overruns, and its value does not exceed the expense.

The next major failures come from the lack of an enterprise approach. Individual

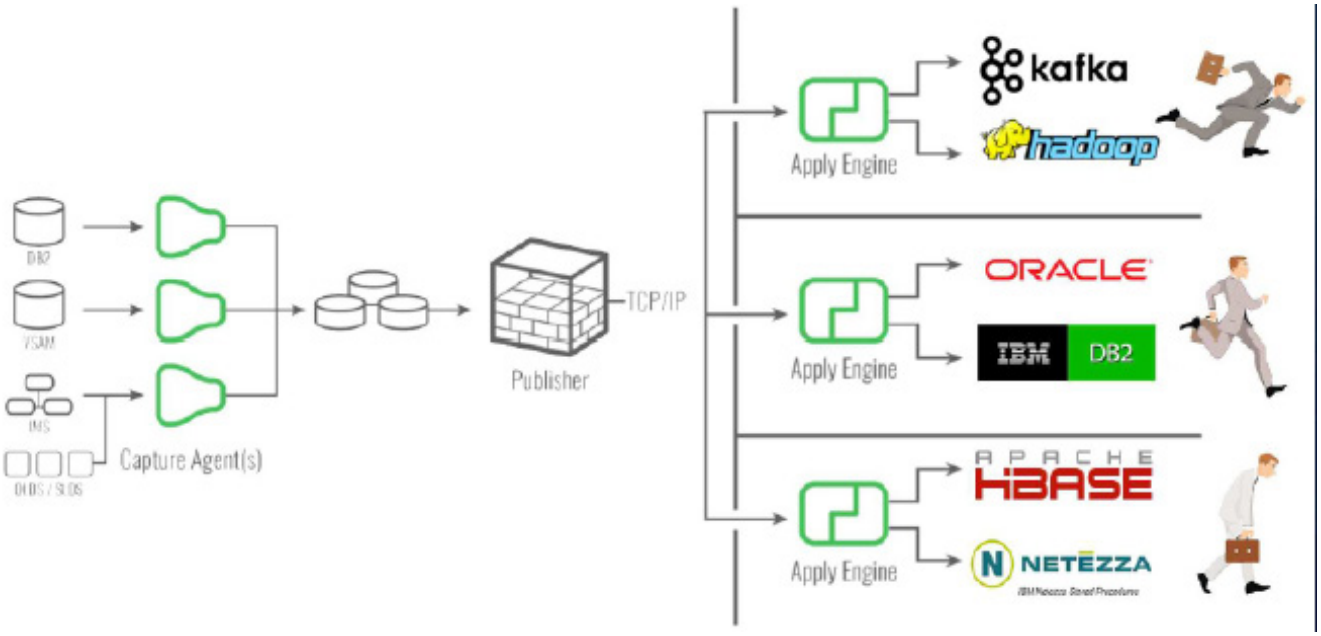



**Figure 4: Target speed and effect on latency**

departments or teams come up with their own different approach. There's minimal overall structure, and it can be more costly to the business.

The last reason for failure is technology. Some sites just copy the data as is into the data lake. They have minimal understanding of the mainframe in general, and non-relational sources pose a significant challenge (eg IMS / VSAM). Mainframe discipline is often lost on Big Data, which results in inappropriate tools being selected that are not aligned with tthe enterprise or not strategic.

Mainframe data streaming is illustrated in Figure 3. Figure 4 shows that the Big Data product chosen can have a big effect on latency.

ETL (Extract, Transform, Load):

- Is used for full data extract / load
- Data transformation logic is defined in this step, then reused by CDC
- Should be run against live data
- Should minimize data landing.

CDC (Changed Data Capture):

- Is used to move only data that has changed
- Re-uses data transformation logic from ETL
- Near-real-time / deferred latency
- Allows for time series analytics.

CDC / ETL data formats include:

- JSON – recommended for data validation
- Avro – recommended for production deployment
- Binary.

When streaming to Hadoop use HDFS format, eg CSV, JSON, Avro. It's typical use is with multiple files with the same content. The file size is based on the number of records / time interval. It requires multi-file management

Partitioning is based on source value(s), eg not native in HDFS, based on source data value(s), and

requires cross-partition multi-file management

Kafka is a high-throughput, low-latency message broker. It was originally written by LinkedIn 2011 and given to Apache in 2012. It supports a variety of targets, with more on the way. It leverages JSON/Avro message format for CDC. It can be used for basic messaging (similar to MQ), Web site activity tracking, metrics collection / monitoring, log aggregation, and streaming.

Scott Quillicy concluded by running through some best practices:

- Approach with a comprehensive strategy – common infrastructure / tools / support; established methods (DevOps / Agile); beware the 'fiefdoms'.
- Involve the business from the beginning – they understand the source data; they know the order of importance; they can assist in design validation, QA, etc.
- Avoid the data collection overkill – time and $$$ killer; focus on most important data first; iterate through remaining data, butprioritize by importance.
- Set proper expectations – 2 to 3 years minimum is expected for an entire project; deliver in increments with the most important data first.
- Understand IMS data is 'special' – patience is key; ask for help.

A copy of Scott presentation is available for download from the Virtual IMS user group Web site at www.fundi.com/virtualims/presentations/IMStoKafkaandHadoopAug17.pdf.

You can see and hear the whole user group meeting by downloading the WMV file from www.fundi.com/virtualims/presentations/2017-08-22meeting.wmv.

## Meeting dates

- On 10 October 2017, Kieco's Henry Kiesslich will be discussing "IMS catalog".
- The following meeting will be on 5 December 2017, when BMC Software will be speaking.

## Recent IMS articles

*Access IMS Data using Node.js – IMS Summer Interns Take the Challenge* by Sandy Sherrill on z Systems Developer Community (11 August 2017). You can find the article at https://developer.ibm.com/zsystems/2017/08/11/ims-summer-interns/.

*Could you use migration assistance, security health check, disaster recovery services, or more?* by Sandy Sherrill on z Systems Developer Community (3 August 2017). https://developer.ibm.com/zsystems/2017/08/03/use-migration-assistance-security-health-check-disaster-recovery-services/

*Use Machine Learning to write HTAP applications for IMS* by Richard Tran on z Systems Developer Community (28 July 2017). You can find the article at https://developer.ibm.com/zsystems/2017/07/28/use-machine-learning-write-htap-applications-ims/.

*IMS and the IBM Z14: More Open and Connected Than Ever* by Kyle J Charlet on z Systems Developer Community (25 July 2017). You can find the article at https://developer.ibm.com/zsystems/2017/07/25/ims-ibm-z14-open-connected-ever/.

*IMS and the IBM z14: Bringing you Trusted Digital Experiences* by S Loomis on z Systems Developer Community (18 July 2017). You can find the article at https://developer.ibm.com/zsystems/2017/07/18/ims-ibm-z14-bringing-trusted-digital-experiences/.

*Protecting Your Most Critical Data: IMS* by Greg Vance on z Systems Developer Community (18 July 2017). You can find the article at https://developer.ibm.com/zsystems/2017/07/18/protecting-critical-data-ims/.

*Help us build a DevCenter that exceeds your expectations* by Sandy Sherrill on z Systems Developer Community (12 July 2017). You can find the article at https://developer.ibm.com/zsystems/2017/07/12/help-us-build-devcenter-exceeds-expectations/.

*Better protection for your IMS data using RACF statistics* by Emily Siu in z Systems Developer Community (11 July 2017). You can find the article at https://developer.ibm.com/zsystems/2017/07/11/better-protection-ims-data-using-racf-statistics/.

*What's the Future for IMS?* by Trevor Eddolls on Destination z (June 2017). You can find the article at http://destinationz.org/Community/Evangelizing-Mainframe/June-2017/What%E2%80%99s-the-Future-for-IMS-.

## Sponsorship opportunity

Are you missing a great opportunity to advertise your IMS software?

The Virtual IMS user group is now offering software vendors the opportunity to advertise their products in a number of ways. You could have an advert on the home page of the Web site (at www.fundi.com/virtualims), you could advertise in the newsletter, and/or you could advertise in the monthly e-mails sent to members of the user group.

E-mail trevor@itech-ed.com for full information about marketing opportunities with the Virtual IMS user group.

**Like us on Facebook**

**#VirtualIMS**

## 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.