# Agenda

1. Brief summary from ISO/IEC JTC 1 Study Group on Big Data – Nancy, Orit, Geoffrey, Arnab, Mark, David, and others

2. Mapping Bob’s ten samples plus other use cases to Ref. Arch. with more detail interface description.

# Action Items

1. Review the breakdown of the scenarios in M0297\_v2, walk through additional ones, and send them around to the reflector.
2. Review the presentations from the ISO meeting

<http://jtc1bigdatasg.nist.gov/programUS.php>

in particular Geoffrey’s

<http://jtc1bigdatasg.nist.gov/_workshop/17_NIST-BigDataOgres.pdf>

for thoughts on the parallels between distributed Big Data Systems, and HPC

# Discussion

Information from the ISO/IEC JTC 1 Study Group on Big Data meeting was discussed. Information from the separate workshop was discussed, as well as the process for submitting information to the study group through INCITS, the U.S. National standards working group looking at Big Data.

The co-located workshop contains presentations from the NIST BDWG, as well as some additional presentations on SQL support for JSON, Software defined networking, data-as-a-service, framework for predicting future technologies, array handling for analytics, high performance computing, characterization of big data systems, big data applications and architectures, security fabric, suggested strategy for standards, and metadata.

It was noted that the purpose of the NIST BDWG is not directly the identification of standards, but the creation of the reports to facilitate communication about Big Data. We do expect as a part of continued Reference Architecture work to identify areas to suggest to the Study Group, but the submission of those ideas will be handled through INCITS.

## Scenarios

The first four scenarios from M0297\_v2 were discussed, to add in more detail to clarify where the different components reside.

**High Level Use Cases Scenario Mapped to Reference Architecture including Opportunities for Standards**

1. **Multiple users performing interactive queries and updates on databases with basic availability**

*Mapping:* Actor is Data Consumer and Data Providers. Framework Providers manage the databases. Application Providers define schemas for the databases and create the application. Users interact with the databases through the Application Provider .

*Reference Architecture Blocks*:

* Actor is both DC and DP
* AP – query interface, rendering, possible summarization?
* AP <-> FP passing query, returning results
* FP data repository, query execution, possible summarization?

*Big Data Issues*:

* distributed SQL,
* consistency across data nodes,
* response times,
* underlying libraries for distributed data access,
* resulting dataset size (summarization)

*Opportunities for standards:* distributed database queries (and SQL queries to Hadoop)

1. **Perform real-time analytics on data streams and notify users when specified events occur**

*Mapping*:Data Providers supply data streams. Application Provider runs code to process data streams, detect events, and notifies DC of events. Framework Providers supplies resources to run the code.

*Reference Architecture Blocks*:

* SO contracts with DC for alerts, systems requirements
* DC specifies alerting condition
* DP supplies data
* AP runs the parsing/fusing and alerting application
* FP platform – real-time queues
* FP structured tree comparison capability
* FP provides infrastructure, runs processing balanced across nodes

*Big Data Issues*:

* Is there a requirement to retain the data? Data Retention Specification?
* What is the fault tolerance requirements for the processing queues

*Opportunities for standards:* Representation of events or processing-metadata standards (more of an ontology question), messaging can be handled like in NIEM, representation of event criteria (regex, conditional tree, etc)

*Context examples*: Telcos, deep packet inspection

**3. Move data from external data sources into a highly horizontally scalable data store, transform it using highly horizontally scalable processing (e.g. Map-Reduce), and return it to the horizontally scalable data store (ELT)**

*Mapping:* Data Provider is the external data source. Framework Provider supplies horizontally scalable data store. Application Provider supplies transformation used by Framework Provider for horizontally scalable processing and return to data store.

(1) Reference Architecture Blocks:

* DP sending “raw” data
* AP may need to parse (extract)
* FP store in distributed datastore (load)

(2) Reference Architecture Blocks (like creating data mart):

* AP query (extract)
* FP run query across distributed datastore (transform), saves results in a new distributed datastore (load)

*Big Data Issues*: ELT vs ETL; Does AP or FP handle transform processing?

*Opportunities for standards:* Standard descriptions of transformations (e.g. JSON and/or XML to tables) and how they’re implemented across a distributed data store.

Note: the transforms could be their own taxonomy

1. **Perform batch analytics on the data in a highly horizontally scalable data store using highly horizontally scalable processing (e.g Map-Reduce) with a user-friendly interface (e.g. SQL like)**

*Mapping:* Framework Provider supplies the horizontally scalable data store and processing engine. Application Provider supplies the analytic software that forms the basis of the processing. SO selects the appropriate analytics to be performed based on requirements from the DC.

*Big Data Issues*: processing libraries

*Opportunities for standards:* Interface between analytics and data processing. Metadata to support data access and data flow for diverse analytic interfaces. Metamodel like PMML for connecting analytics configuration with data query structure.

Note: Does choice of analytics come from AP or SO?

Where is the query processing run – in the application or in the framework

# Attendees

David Boyd

Pw Carey

Geoffrey Fox

Nancy Grady\* compiled minutes

Orit Levin

Bob Marcus

Stephen McGee

Felix Njeh

Mark Underwood

# Chat Log

(1:20 PM) Bob Marcus (ET-Strategies.com): Have to go. Today is my birthday.

(1:20 PM) Bob Marcus (ET-Strategies.com) disconnected.

(1:20 PM) Pw Carey, Compliance Partners, LLC joined.

(1:22 PM) Pw Carey, Compliance Partners, LLC: Happy Birthday Sir Robert....

(1:23 PM) Nancy Grady (SA)C): Study Group

(1:23 PM) Nancy Grady (SA)C): http://jtc1bigdatasg.nist.gov/programUS.php

(1:28 PM) Pw Carey, Compliance Partners, LLC: Did you all discuss; fraud, grc, NSA, privacy, et cetera...?

(1:29 PM) Nancy Grady (SA)C): Look at the presentations on the second day by Fox, Underwood, Melton

(1:29 PM) Mark Underwood (Krypton) disconnected.

(1:29 PM) Nancy Grady (SA)C): Also from Marcus

(1:30 PM) Pw Carey, Compliance Partners, LLC: Ok will do...and thanks....we forgot forensics within a Big Data/Cloud-Ecosystem....?

(1:31 PM) Pw Carey, Compliance Partners, LLC: We think that's their plan.....

(1:31 PM) Pw Carey, Compliance Partners, LLC: No, honest....they look to NIST for direction in these areas.....

(1:34 PM) Pw Carey, Compliance Partners, LLC: Were any personal trust relationships established....?

(1:36 PM) Pw Carey, Compliance Partners, LLC: Did anyone from ISO tell you what were their expectations...?

(1:36 PM) Pw Carey, Compliance Partners, LLC: Going in.....what they hoped to accomplish....?

(1:40 PM) Geoffrey Fox: ISO wants a document suggesting standards areas. Its not clear to me how much detail needed as standards will come from other ISO groups. So we need enough detail to a) Make a case standard needed b) Decide where it should be created

(1:46 PM) Pw Carey, Compliance Partners, LLC: Ok thanks.....&...we agree with your assessment, too

(1:48 PM) Felix Njeh (COMINT) joined.

(1:54 PM) Felix Njeh (COMINT)11 joined.

(1:54 PM) Felix Njeh (COMINT)11 disconnected.

(1:56 PM) John Rogers, HP disconnected.

(1:57 PM) Pw Carey, Compliance Partners, LLC: Ok...we'll review the 297 document and align it with the Cloud Reference Architecture WG efforts.....

(1:58 PM) Geoffrey Fox: sorry I need to go Geoffrey

(1:58 PM) Geoffrey Fox disconnected.

(1:58 PM) Pw Carey, Compliance Partners, LLC: Ok...thanks

(2:00 PM) Pw Carey, Compliance Partners, LLC: Would this be addressed within the confines of the SLA between the Consumer and the Provider...?

(2:01 PM) Pw Carey, Compliance Partners, LLC: We agree with Orit....

(2:04 PM) Pw Carey, Compliance Partners, LLC: Have we defined the Role of the Orchestrator....?

(2:05 PM) Pw Carey, Compliance Partners, LLC: Data Retention Specification....

(2:06 PM) Dave Vennergrund disconnected.

(2:06 PM) Pw Carey, Compliance Partners, LLC: Theoritically...those bits of floating data would be in one of the other two buckets....no?

(2:07 PM) Pw Carey, Compliance Partners, LLC: Allowing us to pick-up seamlessly from the point of the hick-up....yes?

(2:11 PM) Pw Carey, Compliance Partners, LLC: Well....297 is still in DRAFT mode.....yes?

(2:12 PM) Felix Njeh (COMINT) disconnected.

(2:13 PM) Pw Carey, Compliance Partners, LLC: Orit....can you jot down your question(s) as we would like to track them down.....please....Respectfully yours, Pw

(2:19 PM) Pw Carey, Compliance Partners, LLC: Extract--->Translate--->& Load vs Extract--->Load---> Translate....si.....

(2:22 PM) Pw Carey, Compliance Partners, LLC: His initial effort is a good starting point for deeper discussions.....

(2:28 PM) Pw Carey, Compliance Partners, LLC: Thank you....and we think this is a very sharp bunch of cookies, too.....

(2:32 PM) Pw Carey, Compliance Partners, LLC: What should Nancy capture....?

(2:34 PM) Steven McGee joined.

(2:34 PM) Pw Carey, Compliance Partners, LLC: Yes....a very good start....

(2:35 PM) Mark Underwood (Krypton) joined.

(2:35 PM) Mark Underwood (Krypton) disconnected.