NSSDF 2011

Correlation Using Pair-wise Combinations of Multiple Data Sources and Dimensions at Ultra-Large Scales

THE OVERALL CLASSIFICATION OF THIS BRIEFING IS UNCLASSIFIED

Approved for public release;
distribution is unlimited

Jonathan Larson
Kathleen Lossau
Potomac Fusion, Inc
Austin, TX

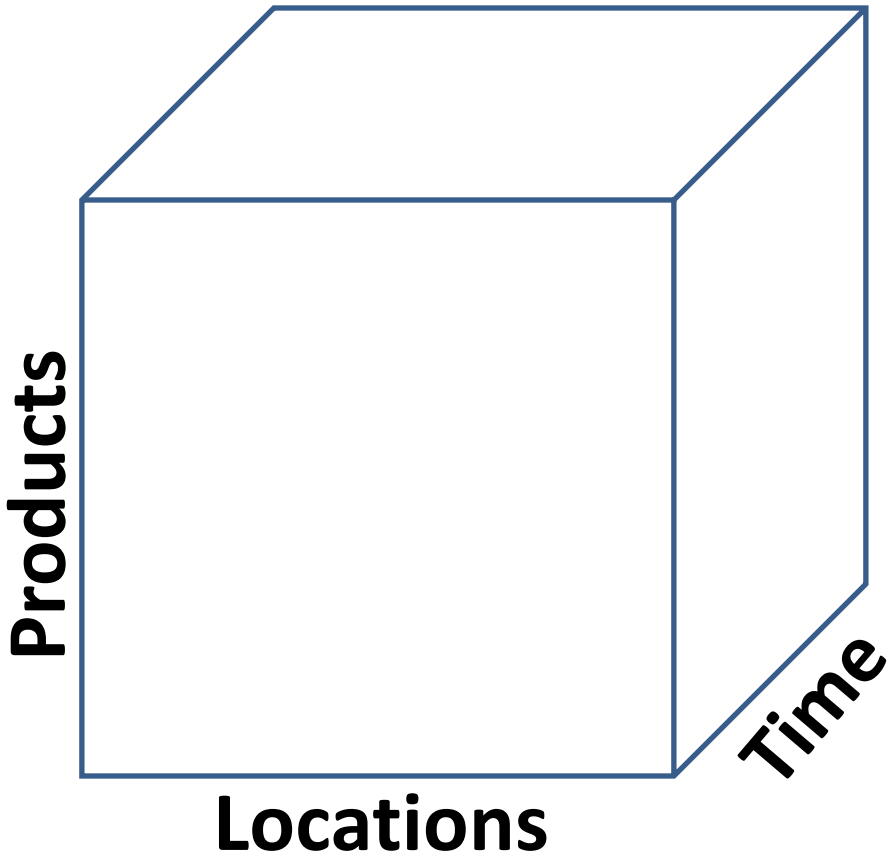
Dale Walsh
MITRE Corporation
McLean, VA

Technical Focus

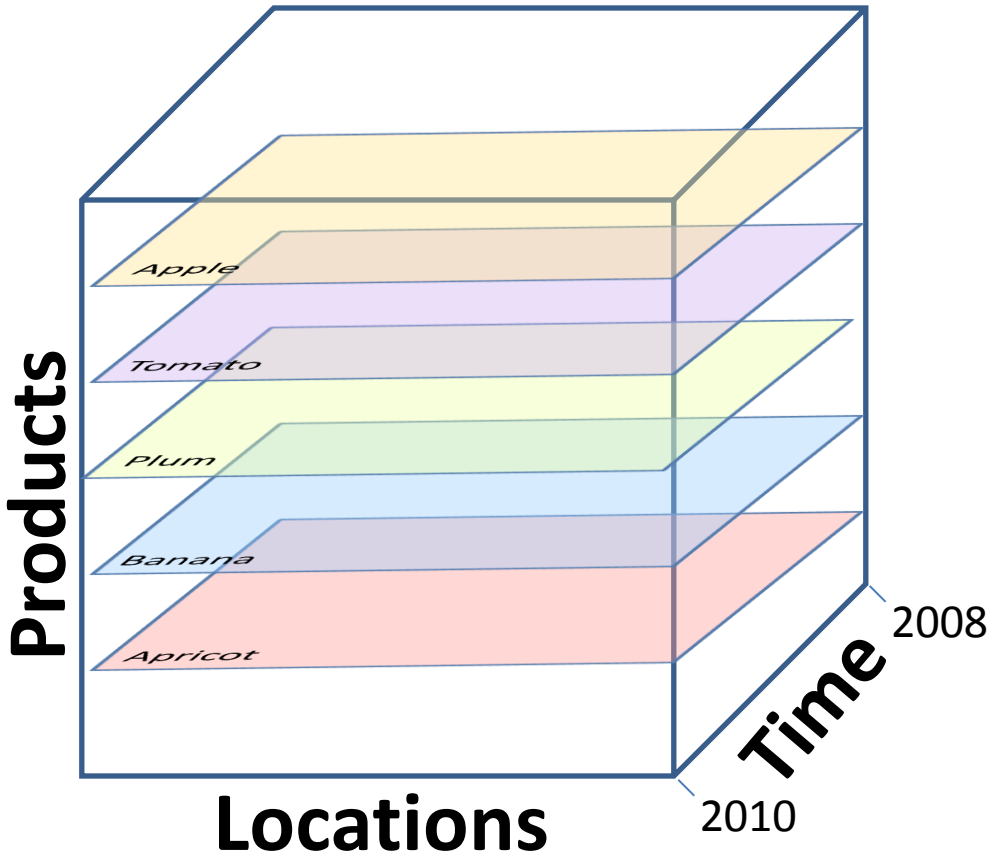
- Current analysis relies heavily on SME knowledge
- Need for automated discovery of non-biased pattern and relationship detectors
- Need for large scale data analytics
  - Traditional single machine data warehousing solutions struggle with scale
  - Data locality requirements cause parallelization challenges
  - Approach to multidimensional database (MDD) implementation in the cloud
  - Using data denormalization techniques to construct planes of enrichment in a distributed environment
- Ability to apply basic analytic techniques at massive scale
  - Supporting massive combinatorialanalysis (ranging into the trillions of comparisons)
  - Assorted distance matrix calculations of feature vectors
  - Covariance / Correlation / Adjacency matrix calculations

Hypercube Introduction

3-D Hypercube

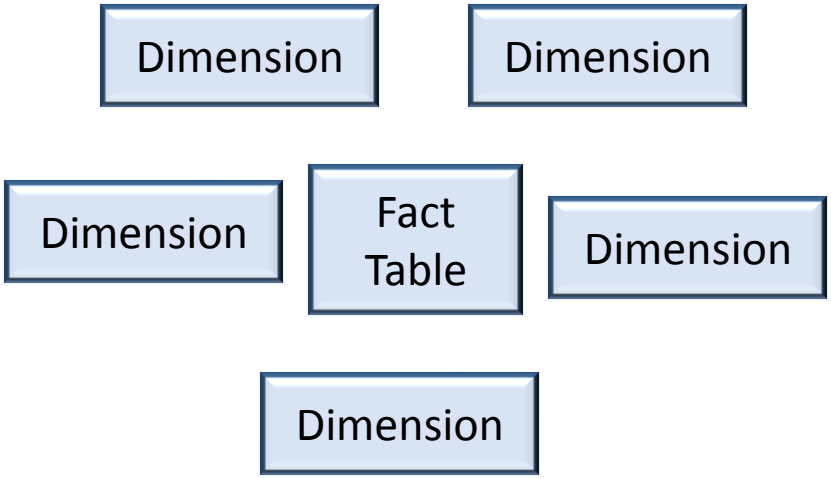


Looking at product layers



MDD Implementations

Example: Star Schema



Example: *1 row of data representing 3 Dimensions "sliced" 9 different ways*

John D	Kabul	2010
John D	2010	Kabul
Kabul	John D	2010
Kabul	2010	John D
2010	Kabul	John D
2010	John D	Kabul

ROLAP — **HOLAP** — **MOLAP**

Normalized
Dynamically Calculated (run time)
Efficient Storage

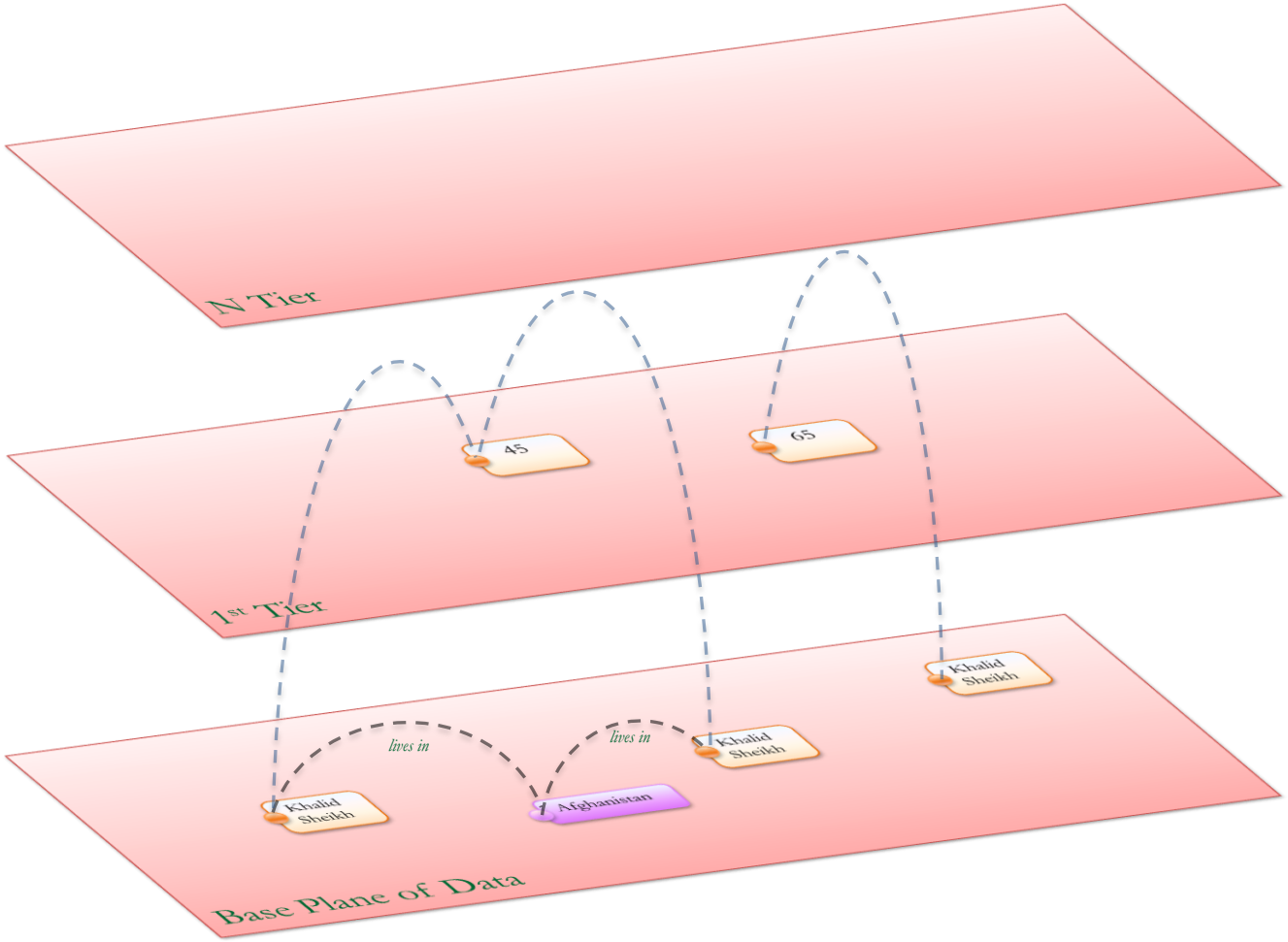
De-normalized
Extremely Performant
Enormous Storage
Specialized Indexes

Hypercube Rollups

- Examples of aggregate rollups
  - Suppose a timestamp of 12:21:06 on January 6, 2011
    - Rollup to Minute / Hour / Day / Month / Week / Day of Week
  - Suppose a latitude / longitude coordinate
    - Rollup to City / District / County / Province / Country
- Aggregation and subsequent serialization greatly expedites derivative analytics – IE: detecting patterns based on day of week on a per province basis. Reduces data footprint for subsequent analytics by orders of magnitude.

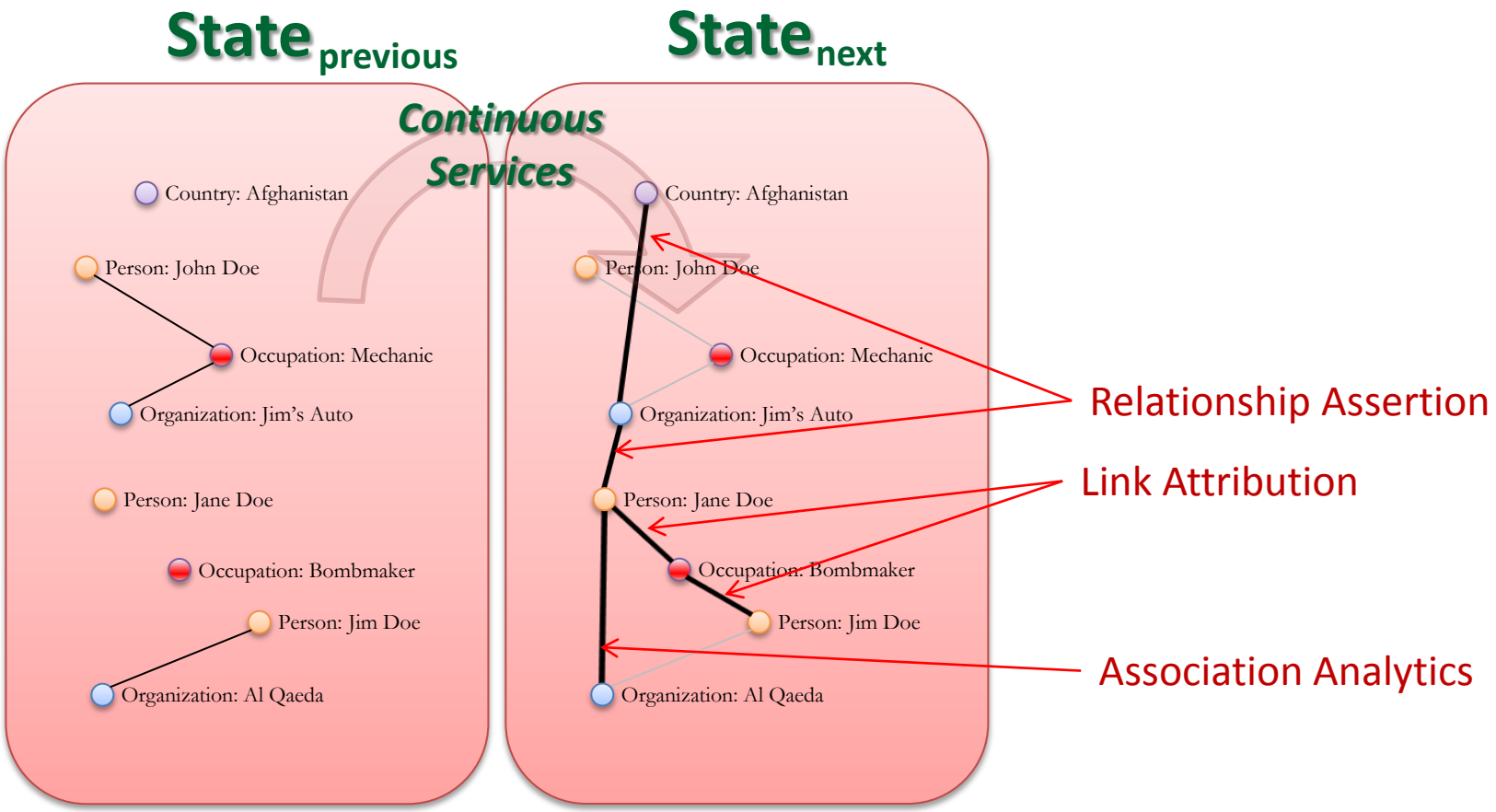
Planes of Enrichment

NSSDF 2011



Planes of Enrichment - Continuous Enrichment

There is no steady state to the system. Each enrichment triggers additional enrichment.



Large Data Analytic Framework

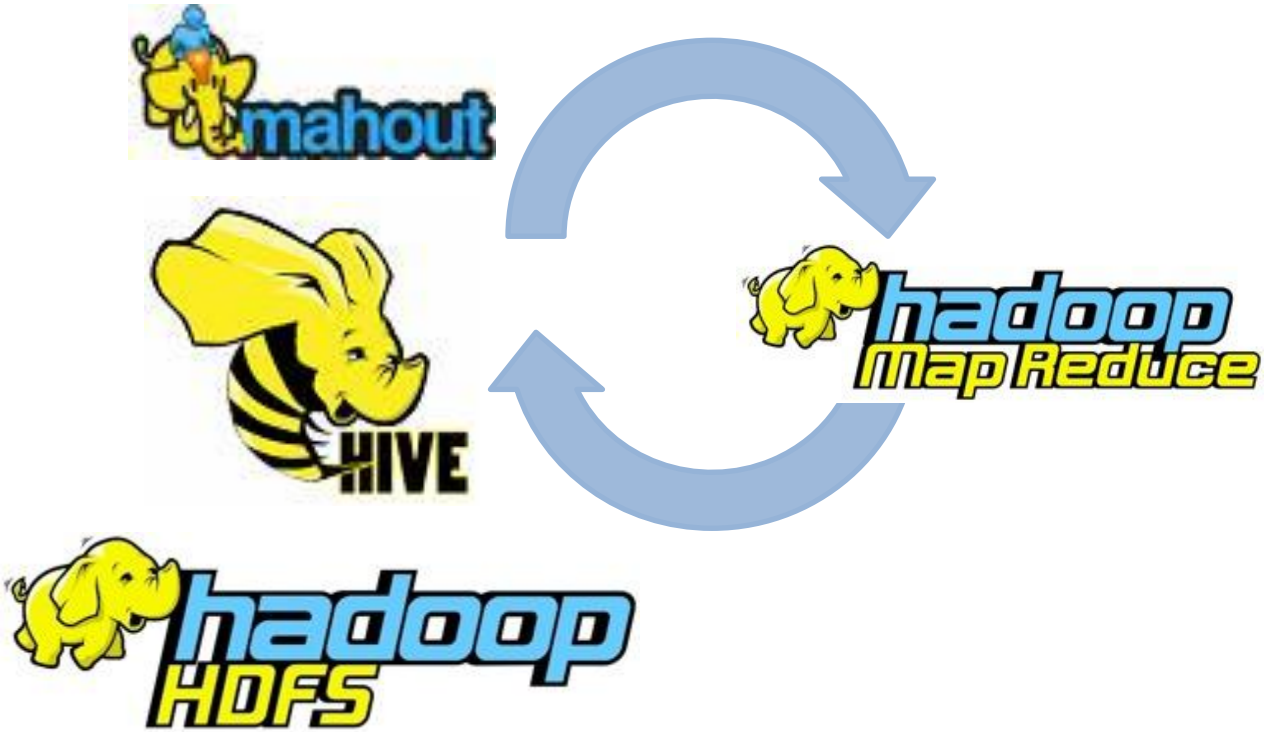
NSSDF 2011

Distributing analytic processing across a cloud of machines using open source technologies

Complex parallelized AI capabilities

Schema information and parallelization of transformations

Raw data and MDD planes of enrichment



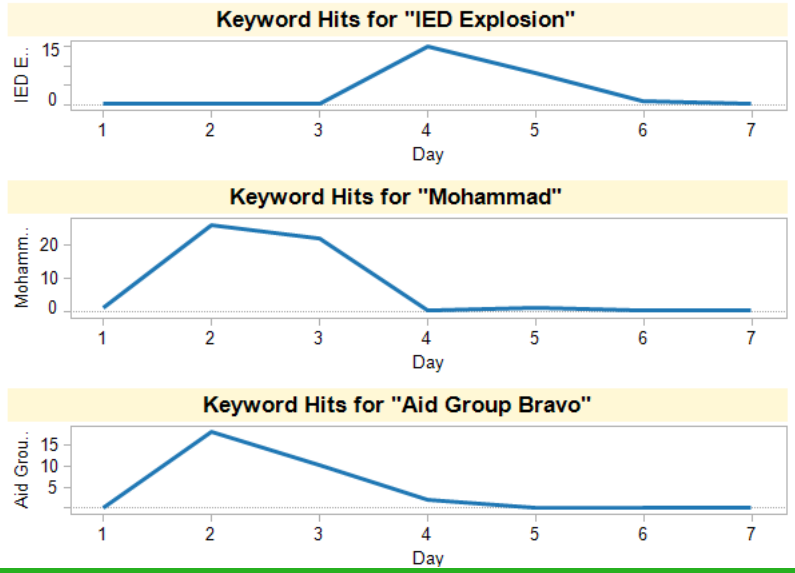
Correlation Analytics

Example correlation scenario:

Alert keyword hit frequency

	Aid Group Bravo	IED Explosion	Mohammad
1	0	0	1
2	18	0	26
3	10	0	22
4	2	15	0
5	0	8	1
6	0	1	0
7	0	0	0

Trend Lines for keyword hits



Pairwise correlation matrix for keyword hit trends

-Correlations > .9 in **bold**

	Aid Group	IED Explosion	Mohammad
Aid Group	**1**	-0.29214	**0.964963**
IED Explosion	-0.29214	**1**	-0.3997
Mohammad	**0.964963**	-0.3997	**1**

Pairwise correlation matrix for keyword hit trends with 2 day lag on y-axis

-Correlations > .9 in **bold**

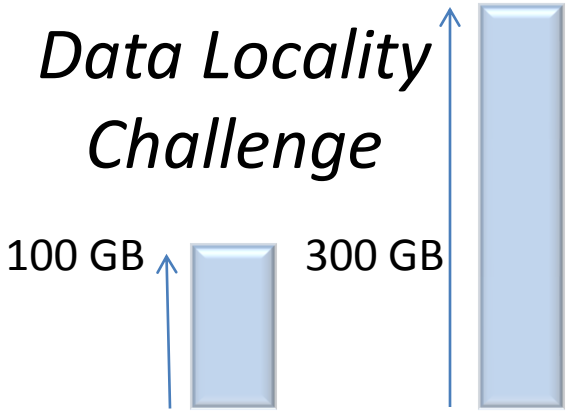
	Aid Group	IED Explosion	Mohammad
Aid Group	-0.09762	-0.31254	-0.09152
IED Explosion	**0.999007**	-0.32393	**0.965013**
Mohammad	-0.25296	-0.27041	-0.20954

Pairwise Combinatorial Analysis

NSSDF 2011

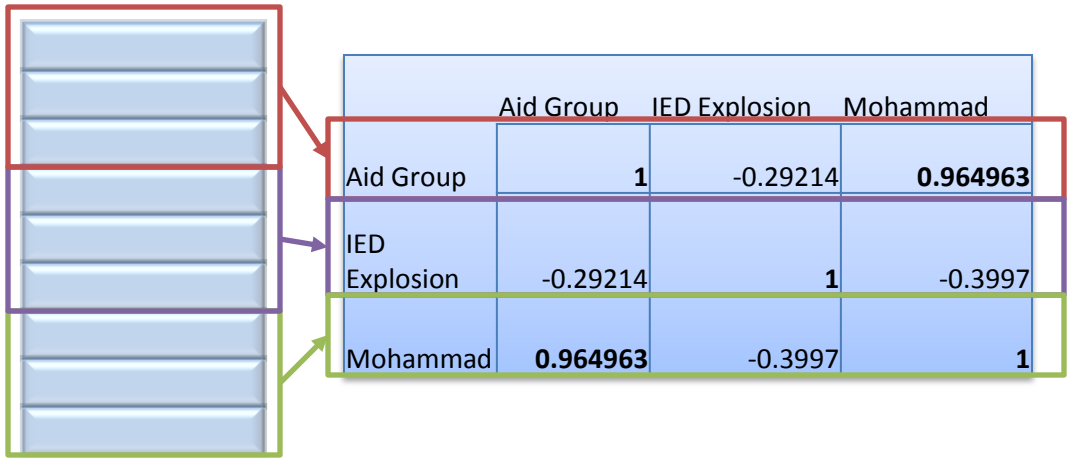
- Combinations build out at ((N^2) – N)/2 pairwise comparisons where N is the number of feature vectors being analyzed.
- For lag window calculations, L*(N^2) pairwise comparisons are needed where N is the number of feature vectors being analyzed and L is the number of lag windows being tested.
- Comparisons ideally require data locality of all data in the same location. Thus, if this exceeds a single machine's RAM specifications, efficient calculation becomes very difficult.
- We have successfully test this technique in a distributed system into the order of trillions of vector value comparisons.

Data Locality Challenge



Memory capacity of single machine

Memory required to calculate correlation matrix



	Aid Group	IED Explosion	Mohammad
Aid Group	**1**	-0.29214	**0.964963**
IED Explosion	-0.29214	**1**	-0.3997
Mohammad	**0.964963**	-0.3997	**1**

Sharding off calculations of correlation matrix

Next Steps

- Investigation into real time processing - Brisk (Cassandra based HDFS)
- Better use of estimation / approximation algorithms (IE: covariance metric estimation)
- Further leveraging cloud based AI packages (IE: Mahout)

Questions?