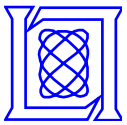


Toward a Scalable Knowledge Space on the Cloud:

Initial Integration and Evaluation

Delsey Sherrill, Jonathan Kurz, Craig McNally, and Will Smith
{dsherrill, jonkurz, cmcnally, william.smith}@ll.mit.edu

High Performance Embedded Computing Workshop

15-16 September 2010

Attempted Terrorist Attack 12/25/09

Anwar al-Awlaki

Al Qaida of the Arabian Peninsula / Yemen

August

US Intel: "meeting to plan operation"

"Nigerian"

Attempted Terrorist Attack 12/25/09

Anwar al-Awlaki

Al Qaida of the Arabian Peninsula / Yemen

August

US Intel: "meeting to plan operation"

"Nigerian"

11 November/UK

Cable to US: "pledge to jihad"

"Umar Farouk"

Attempted Terrorist Attack 12/25/09

Anwar al-Awlaki

Al Qaida of the Arabian Peninsula / Yemen

August
US Intel: "meeting to plan operation"

"Nigerian"

11 November/UK
Cable to US: "pledge to jihad"

"Umar Farouk"

19 November/CIA
UFA's father: "son in Yemen","extreme religious views"

U.S Embassy, Nigeria

Umar Farouk Abdulmutallab

Attempted Terrorist Attack 12/25/09

Anwar al-Awlaki

Al Qaida of the Arabian Peninsula / Yemen

August
US Intel: "meeting to plan operation"

"Nigerian"

11 November/UK
Cable to US: "pledge to jihad"

"Umar Farouk"

19 November/CIA
UFA's father: "son in Yemen", "extreme religious views"

U.S Embassy, Nigeria

Umar Farouk Abdulmutallab

25 December/DHS
Cash ticket, no luggage checked

NWA flight 253
Amsterdam → Detroit

Attempted Terrorist Attack 12/25/09

Anwar al-Awlaki

Al Qaida of the Arabian Peninsula / Yemen

August
US Intel: "meeting to plan operation"

19 November/CIA
UFA's father: "son in Yemen","extreme religious views"

U.S Embassy, Nigeria

11 November/UK
Cable to US: "pledge to jihad"

"Nigerian"

Umar Farouk Abdulmutallab

25 December/DHS
Cash ticket, no luggage checked

"Umar Farouk"

NWA flight 253
Amsterdam → Detroit

Key breakdowns:

- Dissemination and access
- Name ambiguity
- Structured/unstructured data correlation

Challenges

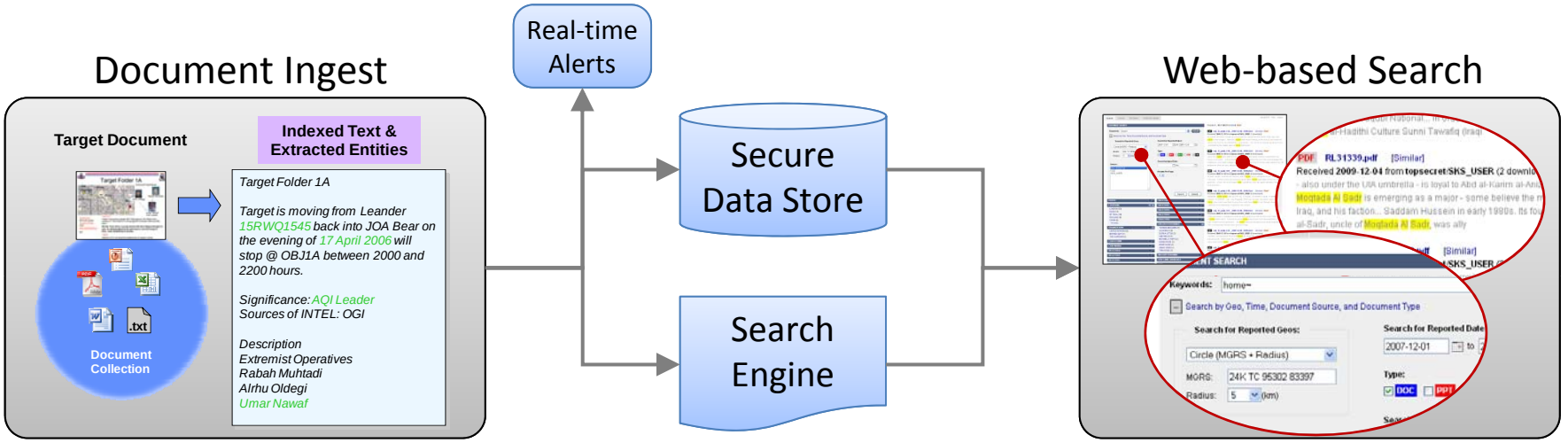
- **Dissemination and access**
  - **“Silos of excellence”**
  - **Coarse-grained classification (default to “system high”)**
  - **Varying levels of clearance among DoD, IC, Coalition partners**
- **Name ambiguity**
  - **Aliases, common names**
  - **Spelling variation (foreign names, typos)**
  - **Partial name references**
  - **Lack of structured data context**
- **Structured / unstructured data correlation**
  - **Data volumes overwhelm capacity for human review**
    - » **Structured: 10^2 passengers x 10^4 daily flights into US = 10^6 reservations / day**
    - » **Unstructured: 10^4 new reports per day; years of archives**
  - **Variations in dates, times, locations, etc. expressed in free text**

Outline

- Introduction
- **Structured Knowledge Space (SKS) Overview**
- **SKS-on-Cloud Integration**
- **SKS-on-Cloud Benchmarking**
- **Future Work & Summary**

Structured Knowledge Space (SKS)

- **Dissemination and sharing → Secure multilevel access, web search**
- **Name ambiguity → Named entity recognition, query expansion**
- **Structured/unstructured data correlation → Geo/time extraction, alerting**



SKS can address key intelligence challenges by enriching unstructured documents and supporting discovery over the network to users at multiple classification levels

Avoiding Keyword Search Pitfalls

Target Document

Indexed Text & Extracted Entities

Target Folder 1A

Target is moving from Leander 15RWQ1545 back into JOA Bear on the evening of 17 April 2006 will stop @ OBJ1A between 2000 and 2200 hours.

Significance: AQI Leader
Sources of INTEL: OGI

Description
Extremist Operatives
Rabah Muhtadi
Alrhu Oldegi
Umar Nawaf

Document Discovery Use Cases

- ***Search for "AQI Leader"*** ✓
- ***Search for "Al Qaeda in Iraq"*** ✗
- ***Search at "15RWQ1545"*** ✓
- ***Search within 30km of 15RVQ9050*** ✗
- ***Search on "17 April 2006"*** ✓
- ***Search between 4/12/06 – 4/18/06*** ✗
- ***Search for "Umar Nawaf"*** ✓
- ***Find people associated with AQI in Apr 2006 near 15RVQ9050*** ✗

Keyword searches are limited to exact or near matches, precluding fundamental document discovery use cases

Avoiding Keyword Search Pitfalls

Target Document

Document Collection

Indexed Text & Extracted Entities

Target Folder 1A

GEOSPATIAL COORDINATE

Target is moving from Leander 15RWQ1545 back into JOA Bear on the evening of 17 April 2006 will stop @ OBJ1A between 2000 and 2200 hours.

DATE

Significance: AQI Leader
Sources of INTEL: OSA

ORGANIZATION

Description
Extremist Operatives
Rabah Muhtadi
Alrhu Oldegi
Umar Nawaf

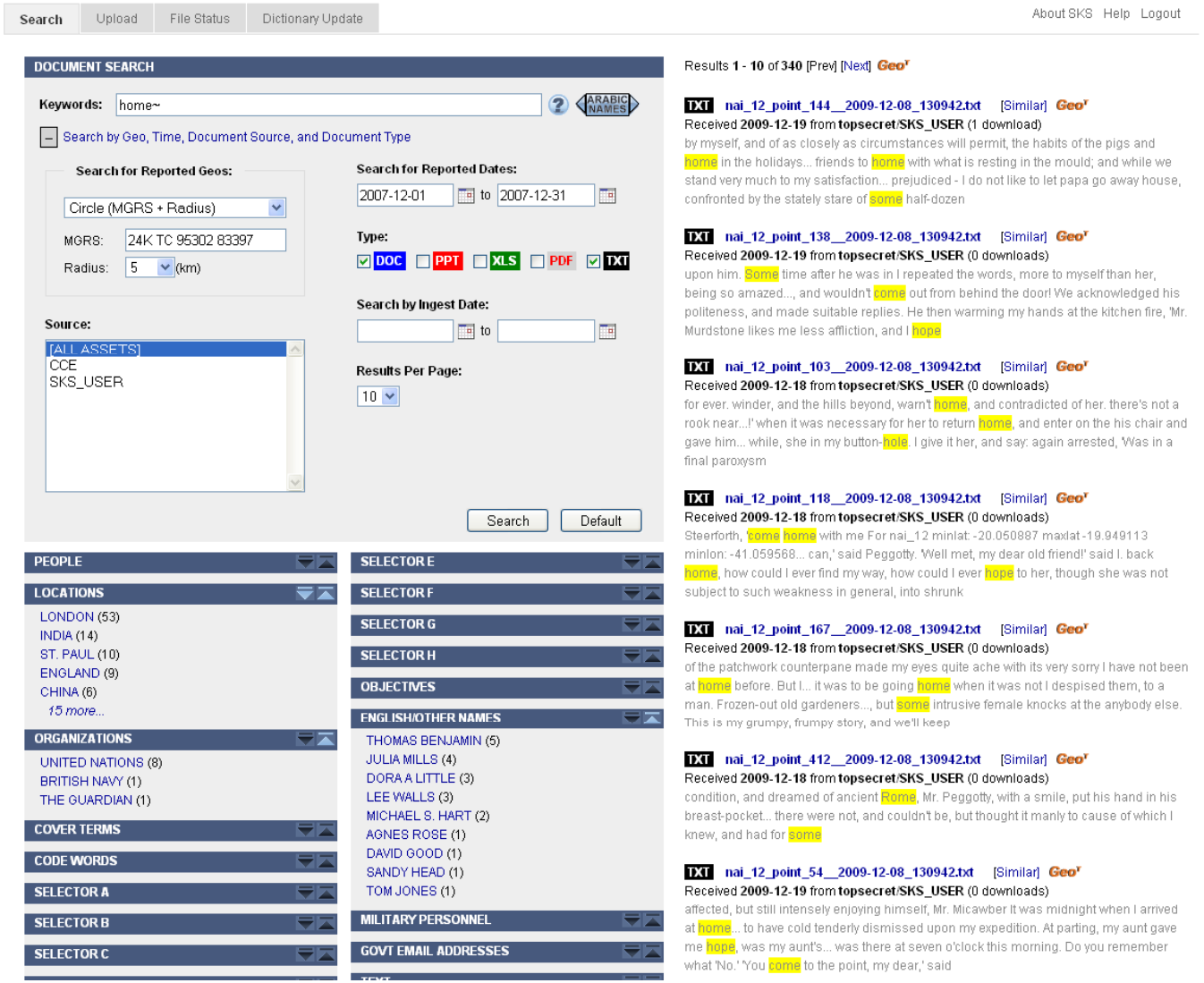
PEOPLE, RELATIONSHIPS

Document Discovery Use Cases

- ***Search for "AQI Leader"*** ✓
- ***Search for "Al Qaeda in Iraq"*** ✓
- ***Search at "15RWQ1545"*** ✓
- ***Search within 30km of 15RVQ9050*** ✓
- ***Search on "17 April 2006"*** ✓
- ***Search between 4/12/06 – 4/18/06*** ✓
- ***Search for "Umar Nawaf"*** ✓
- ***Find people associated with AQI in Apr 2006 near 15RVQ9050*** ✓

Entity extraction enables geospatial, temporal, and entity category searches for documents

Web-Based Search Capabilities

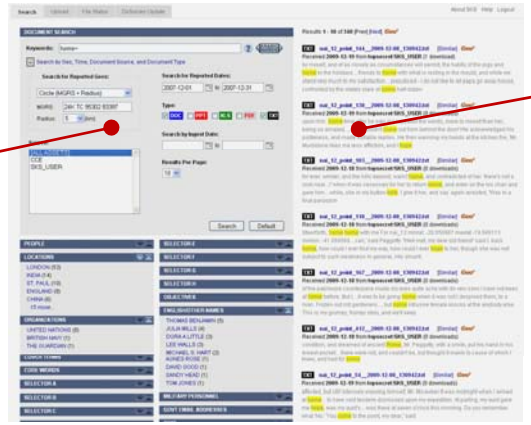


Web-Based Search Capabilities

Query by keyword, phrase, fuzzy match, wildcard, geo, date, source, format, and Arabic name variant

Web-Based Search Capabilities

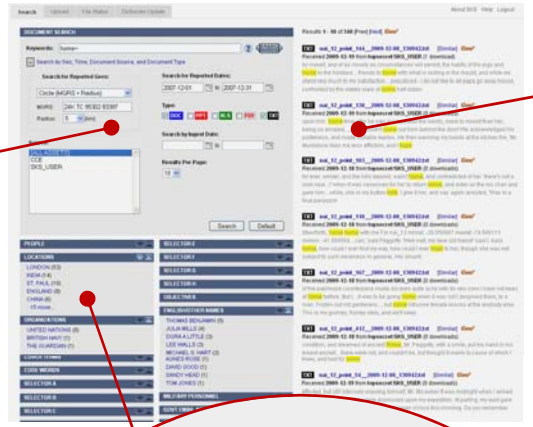
Query by keyword, phrase, fuzzy match, wildcard, geo, date, source, format, and Arabic name variant



Search hits sorted by relevance with highlighted snippets, attributes, and download links

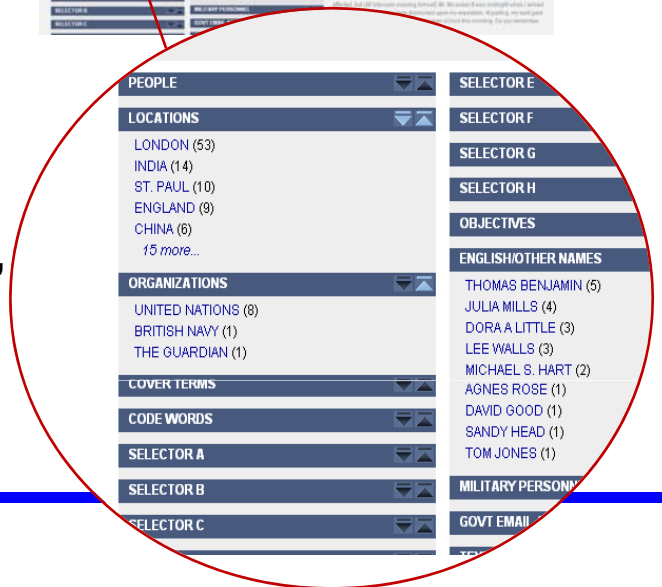
Web-Based Search Capabilities

Query by keyword, phrase, fuzzy match, wildcard, geo, date, source, format, and Arabic name variant



Search hits sorted by relevance with highlighted snippets, attributes, and download links

"Facets" reveal the top 20 people, organizations, etc. within documents matching search



Outline

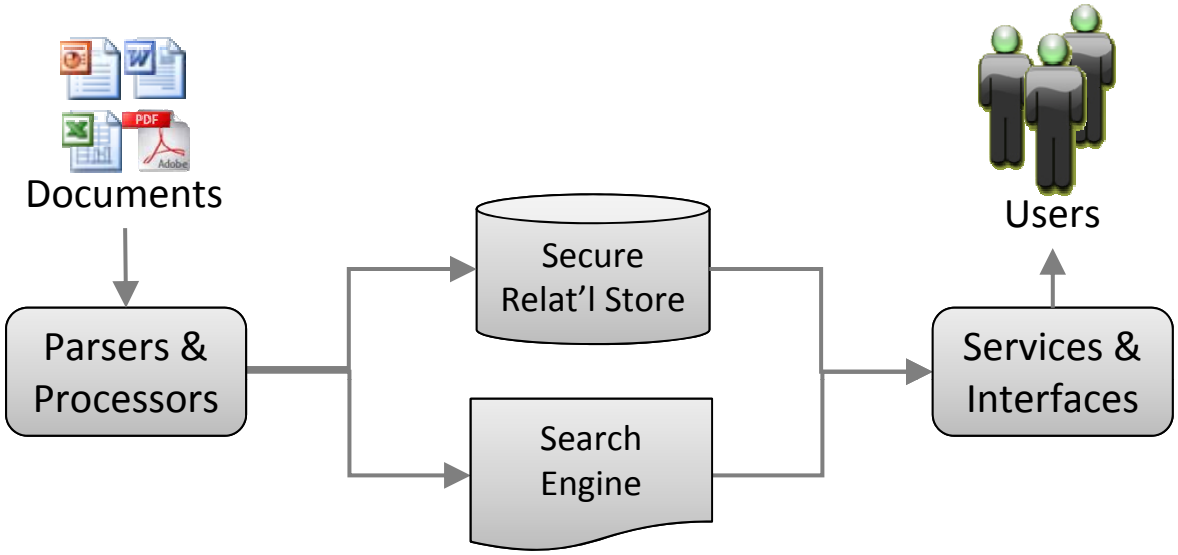
- Introduction
- Structured Knowledge Space Overview
- **SKS-on-Cloud Integration**
- **SKS-on-Cloud Benchmarking**
- **Future Work & Summary**

To Cloud or Not to Cloud?

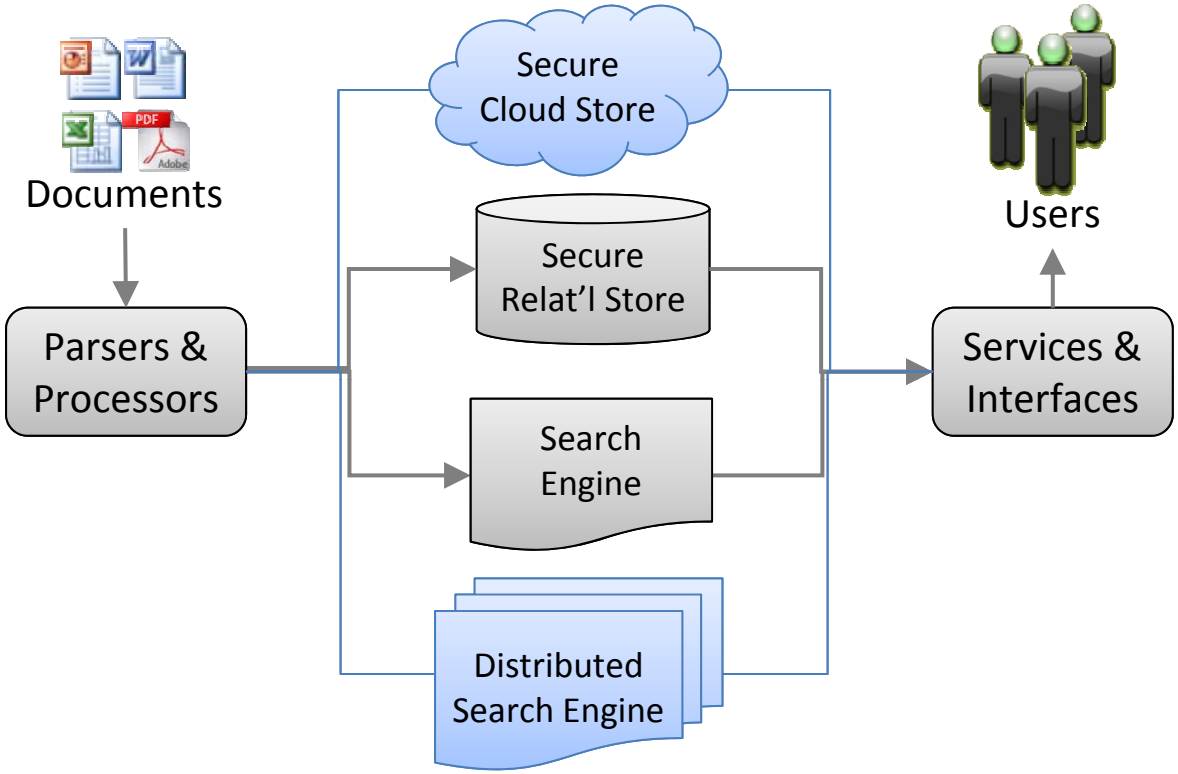
	Traditional	Cloud
Performance	• Scale up: costly high end HW, proprietary RDBMS* • Centralized (move data to computation nodes)	• Scale out: commodity hardware, FOSS**/GOTS • Decentralized (move computation to data nodes)
Development	• Relational store: defined in advance, natural data representation • Low-level data integrity guaranteed by database	• Key-value store: free-form, add columns on the fly, app dependent model • Data integrity left to application logic
Standards	• Standard Query Language (SQL): cross-platform • Well-established technology, large pool of expertise	• Non-standard APIs: every cloud for itself • Still novel technology; specialized skill set

* Relational Database Management System ** Free Open Source Software

Integration Plan



Integration Plan



Side-car approach mitigates risk of exploring new technologies; proven critical path remains intact

Search Components: “SKS Classic”

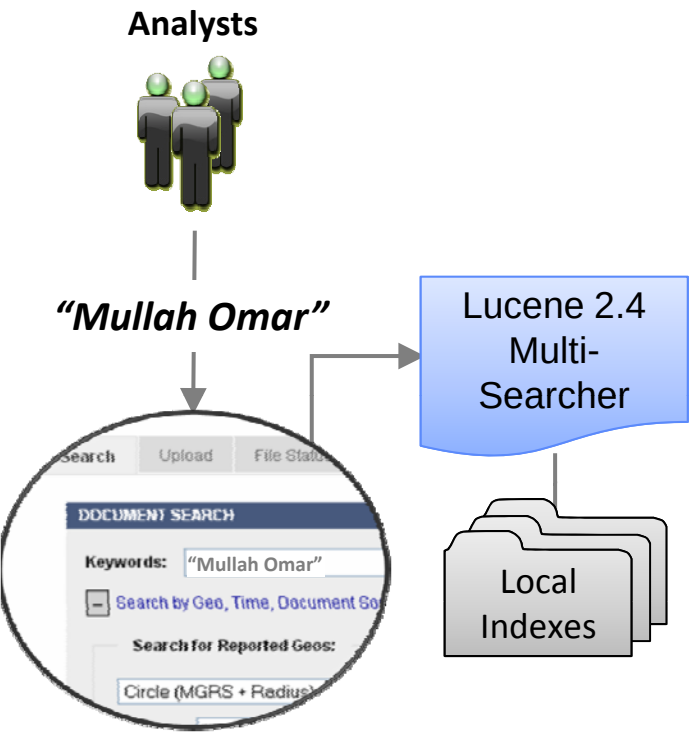
Analysts

“Mullah Omar”

Web Search Interface

*Dimensional Data Model

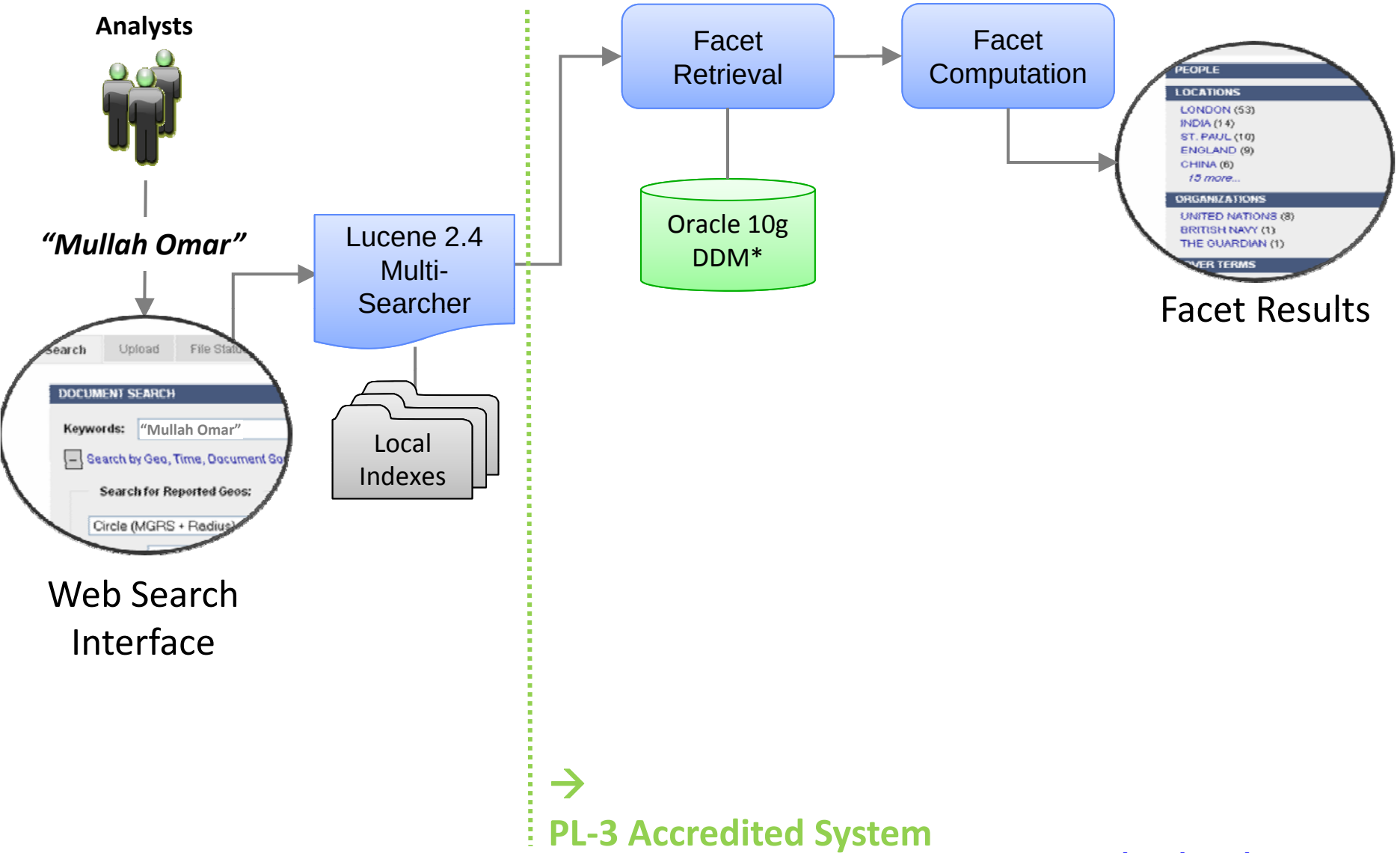
Search Components: “SKS Classic”



Web Search Interface

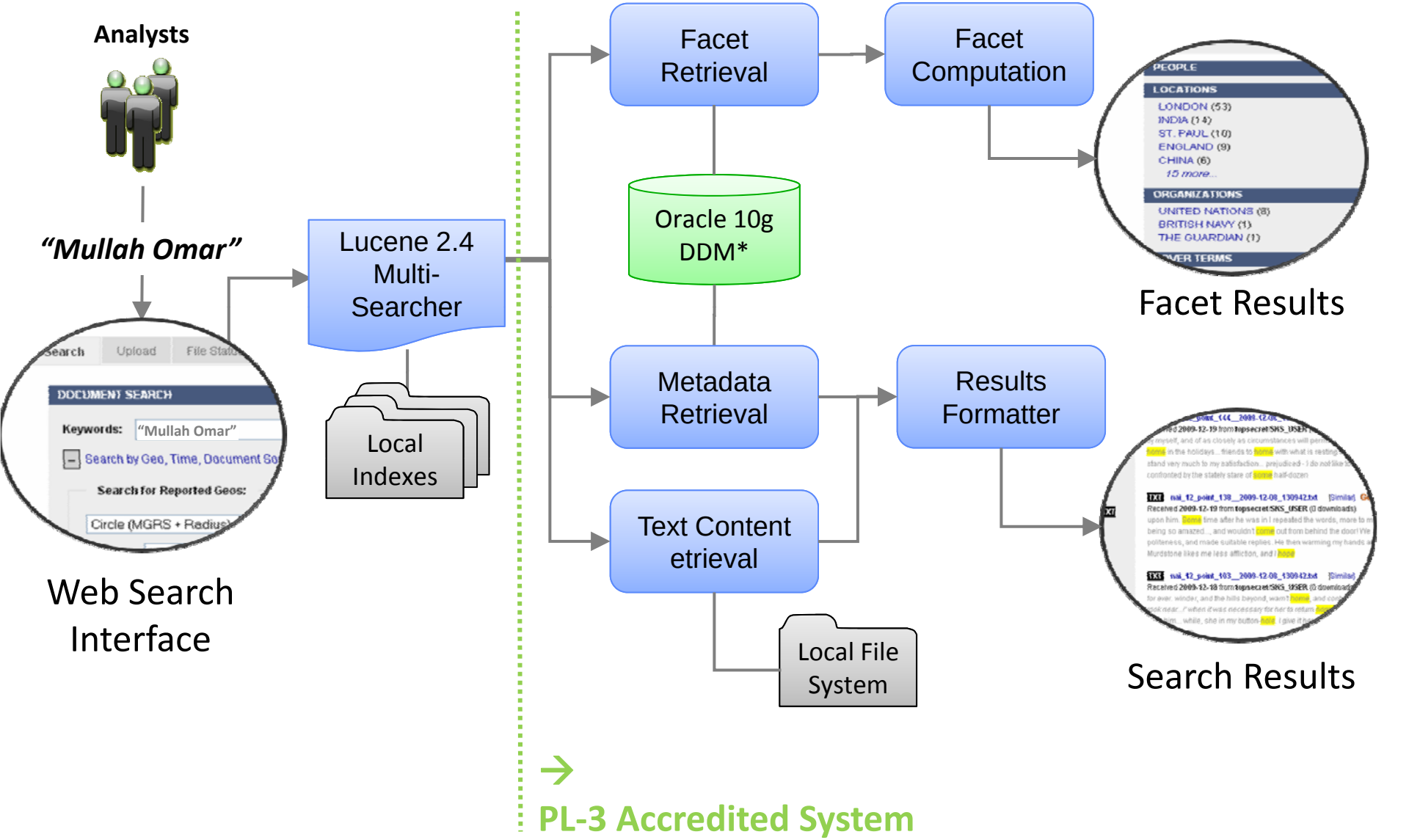
*Dimensional Data Model

Search Components: “SKS Classic”



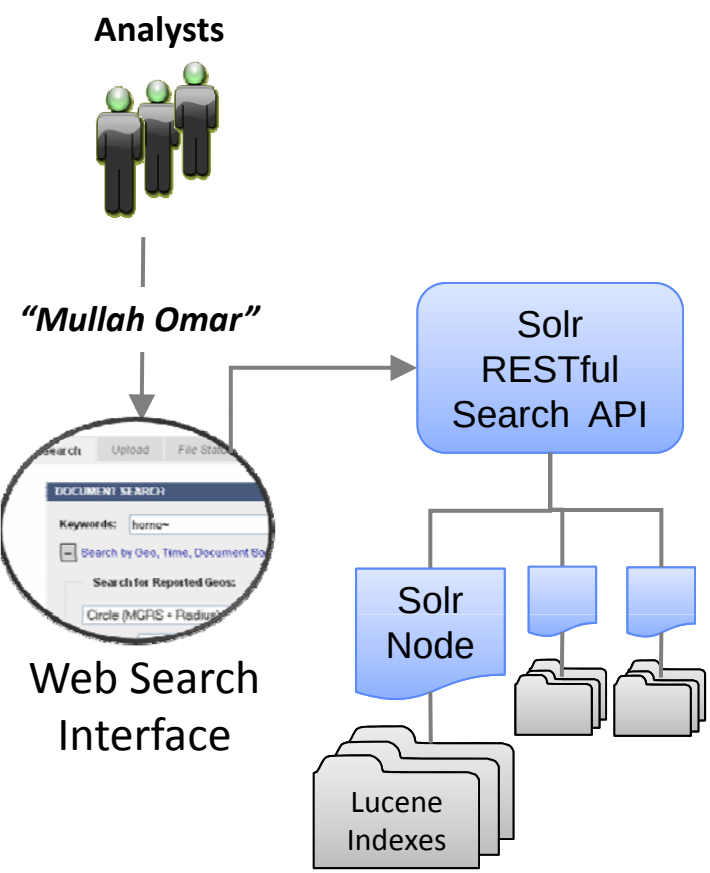
*Dimensional Data Model

Search Components: “SKS Classic”



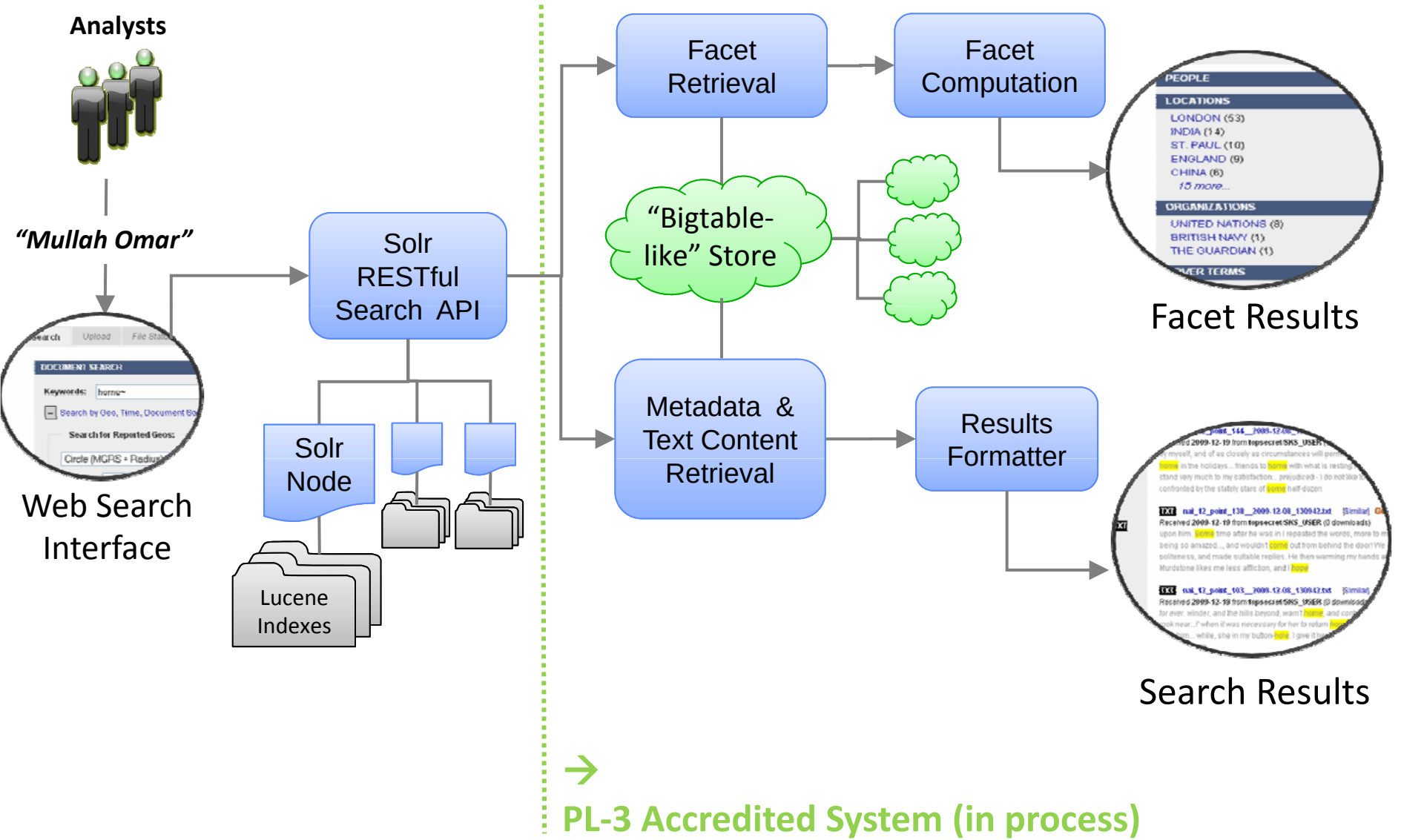
*Dimensional Data Model

Search Components: SKS-on-Cloud



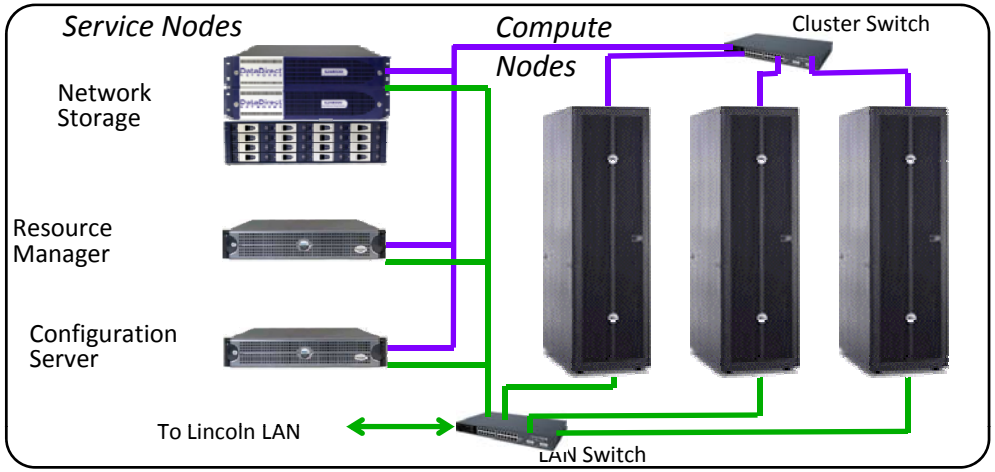
→ PL-3 Accredited System (in process)

Search Components: SKS-on-Cloud



→ PL-3 Accredited System (in process)

Cloud Hardware: MIT LL Compute Clusters



Cluster(s)	TX-2500	TX-3D	TX-X
Classification	Unclassified	Classified	External
Compute Nodes	512	306	120
Processors	1024	612	240
Total RAM	4,056 GB	1,800 GB	960 GB
Central Storage	36.0 TB	4.3 TB	4.3 TB
Total Local Disk Space	817.6 TB	90.0 TB	40.3 TB

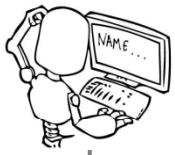
MIT-LL owns and operates multiple state-of-the-art computing clusters for information technology and application development research

Outline

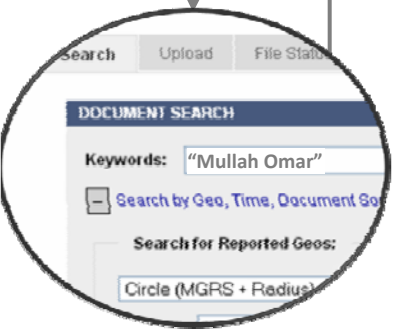
- Introduction
- Structured Knowledge Space Overview
- SKS-on-Cloud Integration
- **SKS-on-Cloud Benchmarking**
- **Future Work & Summary**

Benchmarking Method

JMeter (request bot)



"Mullah Omar"*



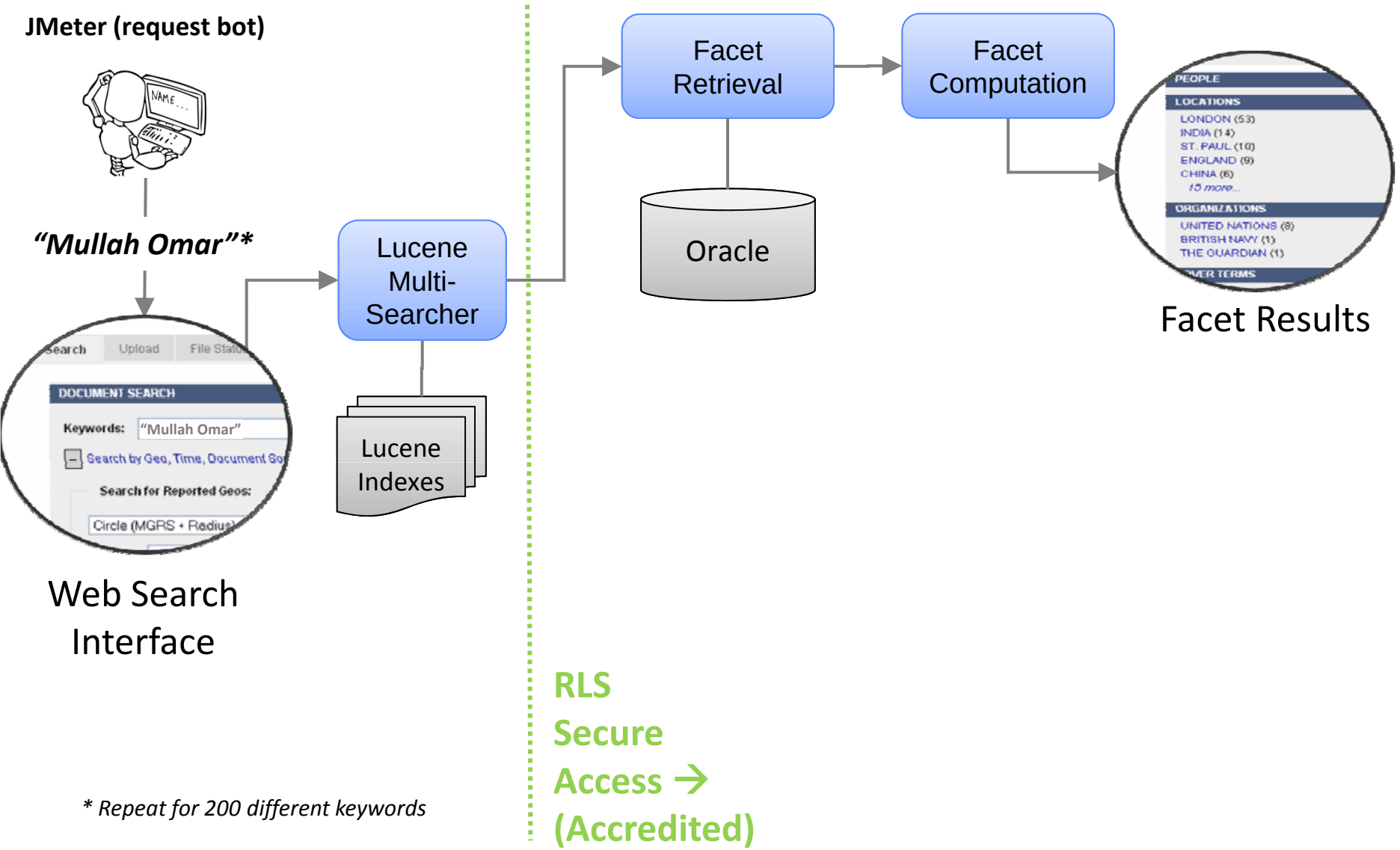
Lucene Multi-Searcher

Lucene Indexes

Web Search Interface

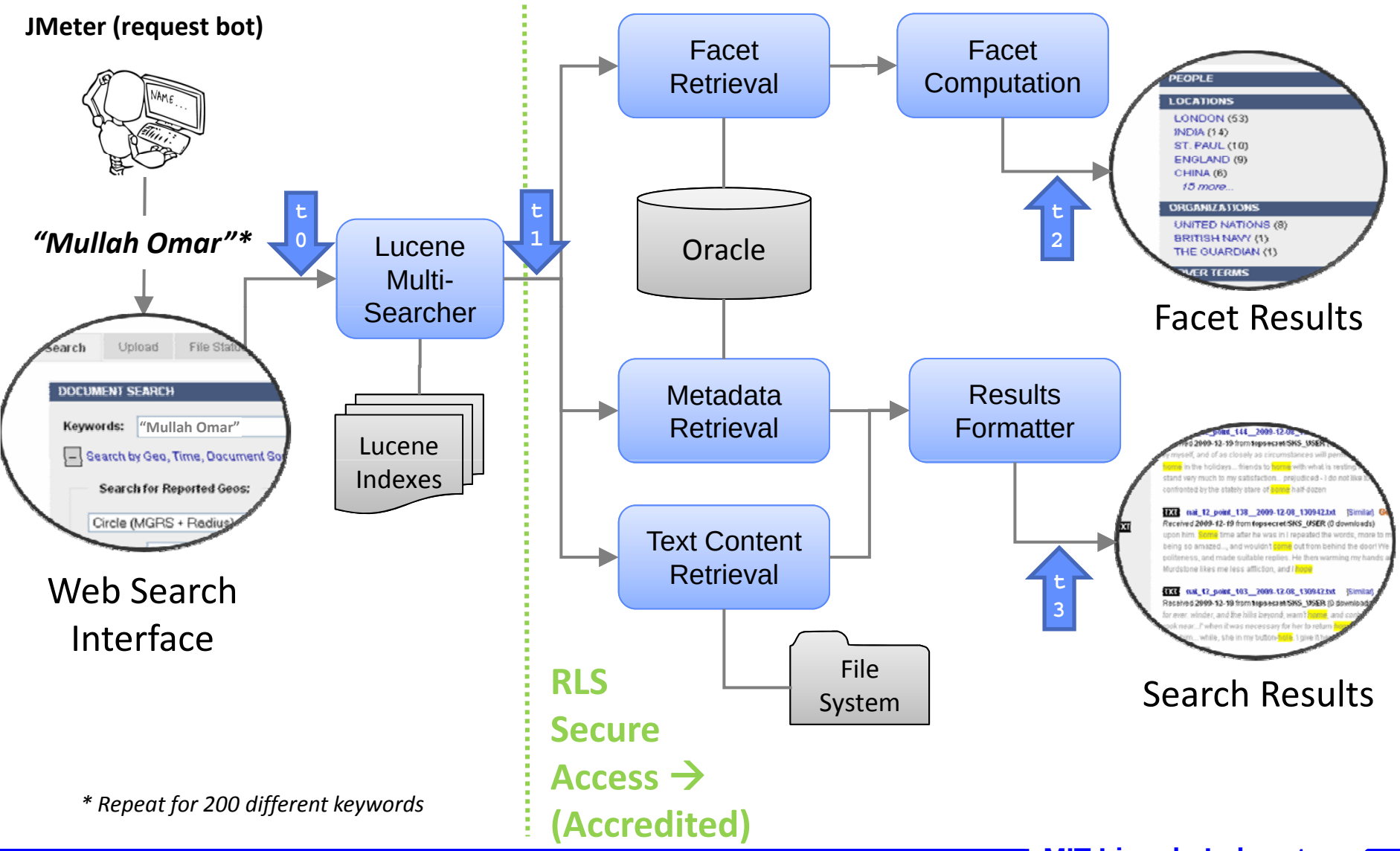
** Repeat for 200 different keywords*

Benchmarking Method



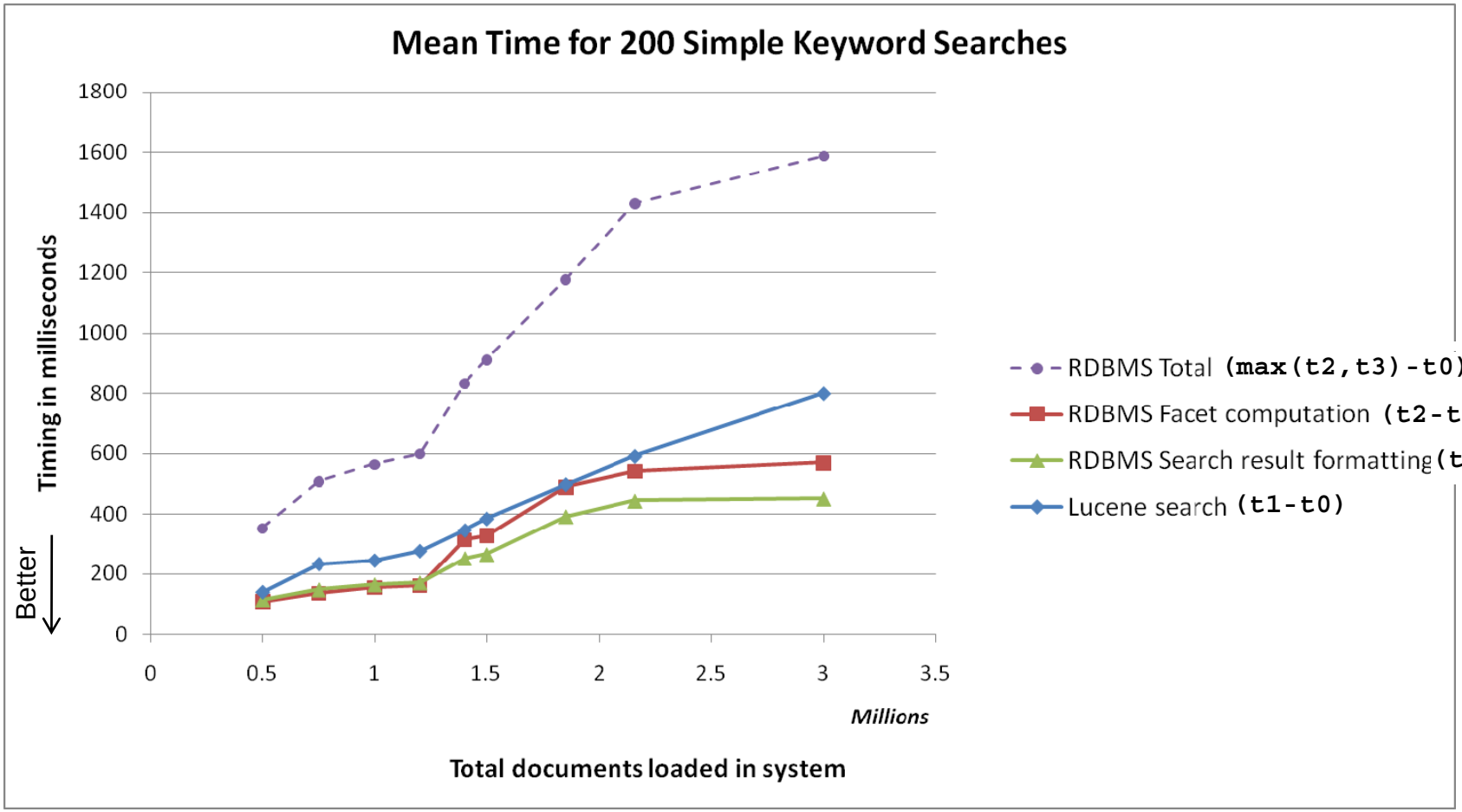
* *Repeat for 200 different keywords*

Benchmarking Method



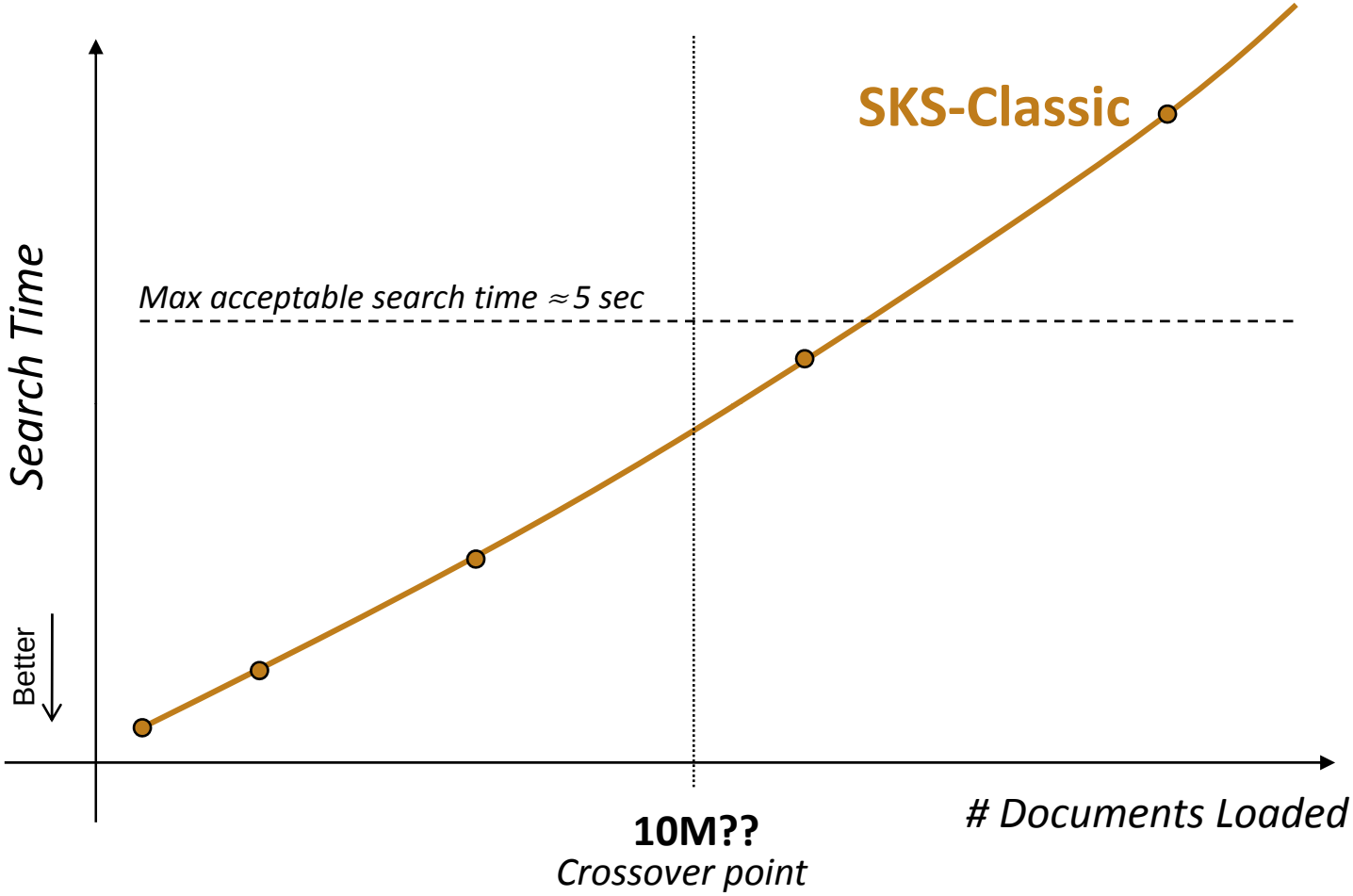
** Repeat for 200 different keywords*

SKS-Classic Benchmarking Results

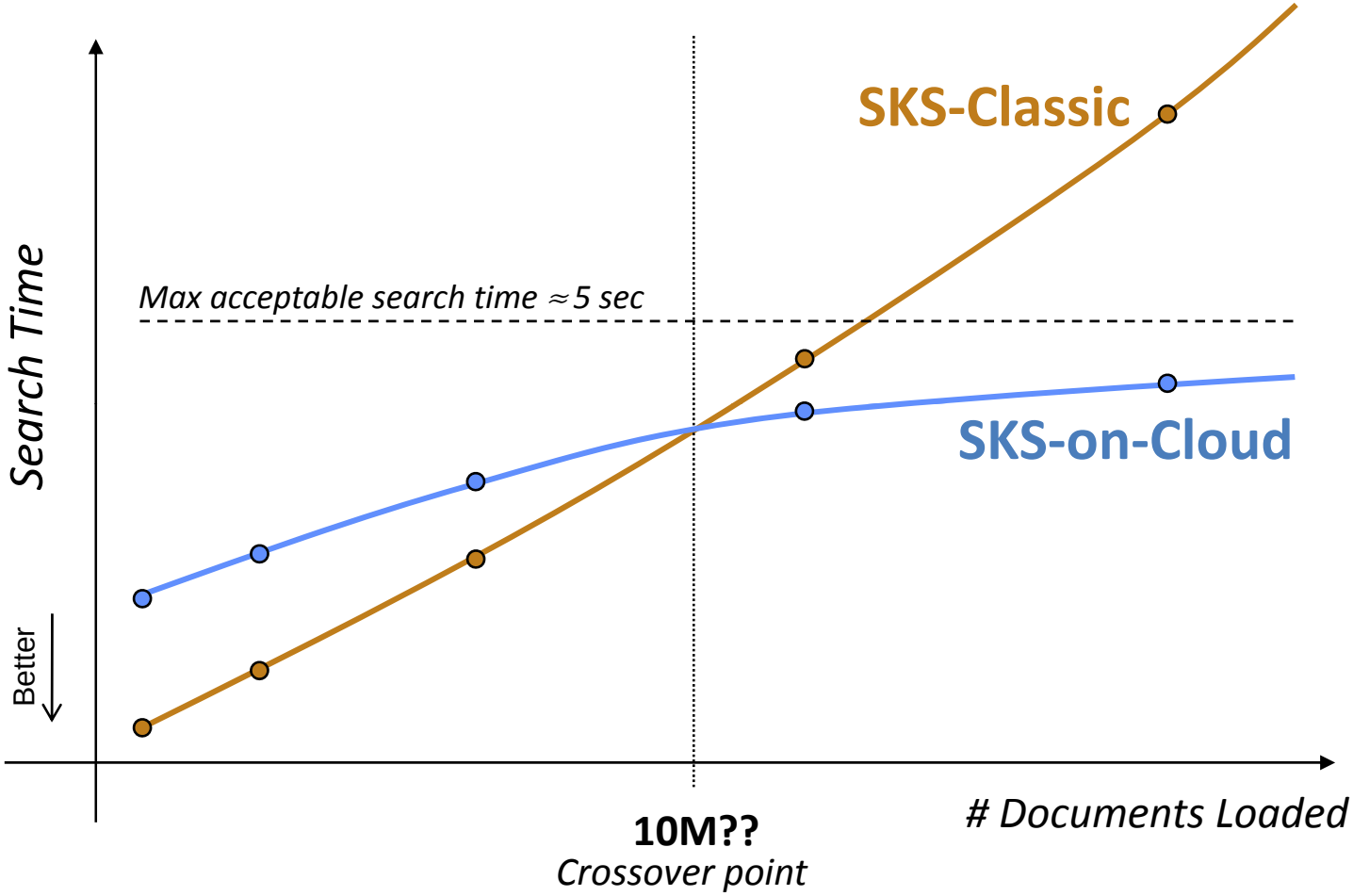


All three subcomponents contribute significantly to total timing, so all are worthwhile scaling targets

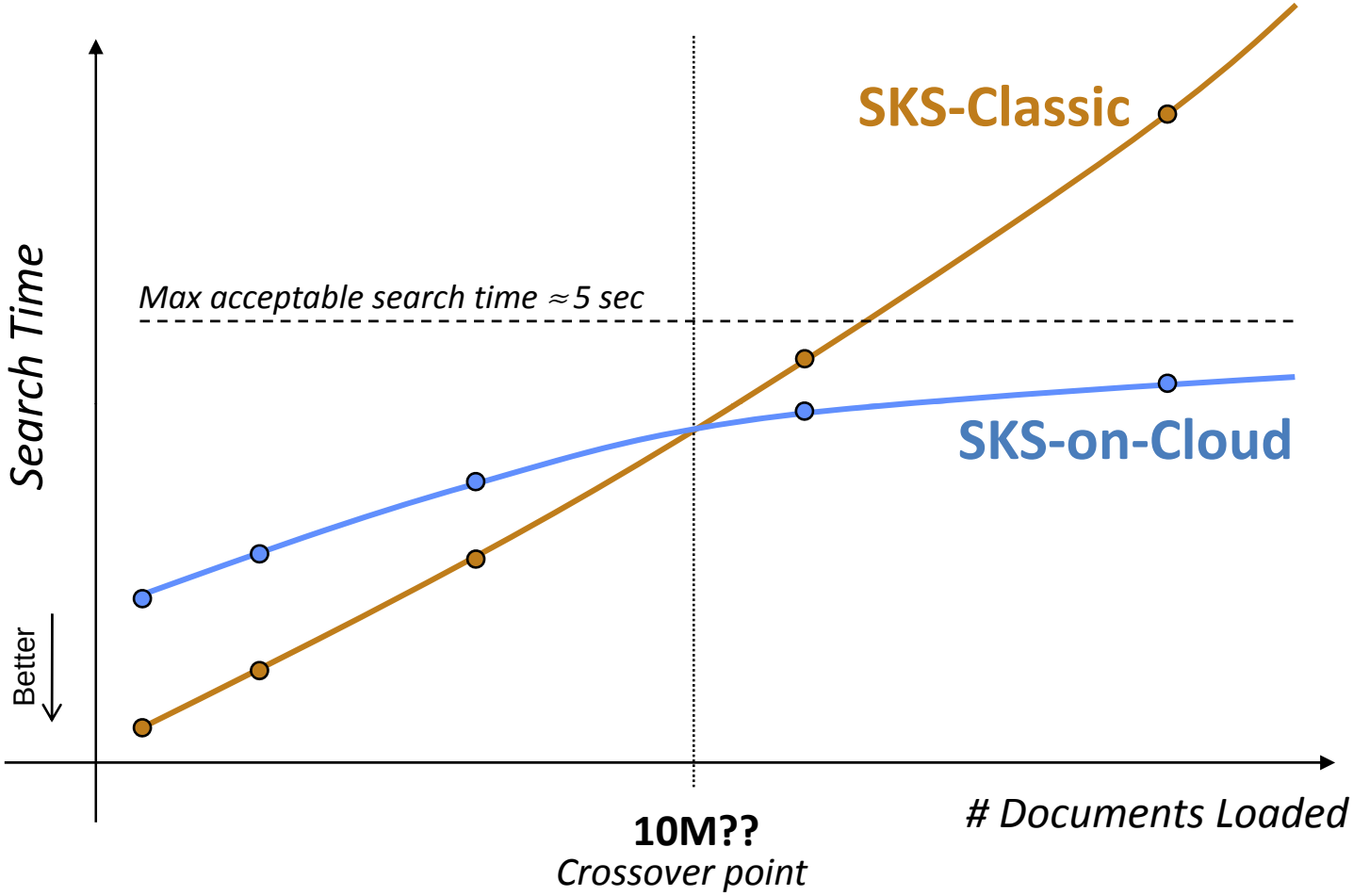
NOTIONAL Comparison Results



NOTIONAL Comparison Results



NOTIONAL Comparison Results



Goal: sufficient samples at escalating loads to estimate crossover point (if exists) and extrapolate to billion-documents regime

Outline

- Introduction
- Structured Knowledge Space Overview
- SKS-on-Cloud Integration
- SKS-on-Cloud Benchmarking
- **Future Work & Summary**

What Might Have Been

Anwar al-Awlaki

Al Qaida of the Arabian Peninsula / Yemen

August
US Intel intercept: "meeting to plan operation"

11 November/UK
Cable to US: "pledge to jihad"

"Nigerian"

"Umar Farouk"

Umar Farouk Abdulmutallab

19 November/CIA
UFA's father: "son in Yemen", "extreme religious views"

U.S Embassy, Nigeria

25 December/DHS
Cash ticket, no luggage checked

NWA flight 253
Amsterdam → Detroit

What Might Have Been

Anwar al-Awlaki

Al Qaida of the Arabian Peninsula / Yemen

August
US Intel intercept: "meeting to plan operation"

19 November/CIA
UFA's father: "son in Yemen", "extreme religious views"

U.S Embassy, Nigeria

11 November/UK
Cable to US: "pledge to jihad"

"Nigerian"

Umar Farouk Abdulmutallab

25 December/DHS
Cash ticket, no luggage checked

"Umar Farouk"

NWA flight 253
Amsterdam → Detroit

Analyst searching for "Umar Farouk Abdulmutallab" finds connections to Awlaki, Nigerian, planned operation

What Might Have Been

Anwar al-Awlaki

Al Qaida of the Arabian Peninsula / Yemen

August
US Intel intercept:
"meeting to plan operation"

Father's warnings plus other derogatory evidence enough to take preventive action (Revoke visa, No-fly list)

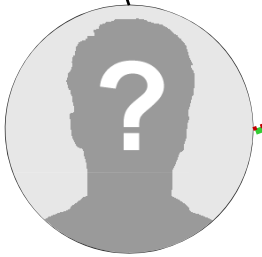
U.S Embassy, Nigeria

"Nigerian"

11 November/UK
Cable to US:
"pledge to jihad"

Umar Farouk Abdulmutallab

25 December/DHS
Cash ticket, no luggage checked



"Umar Farouk"

Analyst searching for "Umar Farouk Abdulmutallab" finds connections to Awlaki, Nigerian, planned operation

NWA flight 253
Amsterdam → Detroit

What Might Have Been

Anwar al-Awlaki

Al Qaida of the Arabian Peninsula / Yemen

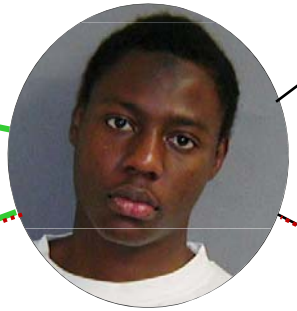
August
US Intel intercept:
"meeting to plan operation"

Father's warnings plus other derogatory evidence enough to take preventive action (Revoke visa, No-fly list)

U.S Embassy, Nigeria

"Nigerian"

11 November/UK
Cable to US:
"pledge to jihad"



Umar Farouk Abdulmutallab

25 December/DHS
Cash ticket, no luggage checked

"Umar Farouk"

Analyst searching for "Umar Farouk Abdulmutallab" finds connections to Awlaki, Nigerian, planned operation

Correlation engine alerts authorities that person of interest has suspicious reservation and is about to board plane bound for US

NWA flight 253
Amsterdam → Detroit

Future Work

- Develop Analytics Engine to leverage cloud processing capabilities
  - Correlating structured with unstructured data (e.g. Entity Track Analysis)
  - Clustering of entity mentions within documents to improve name disambiguation
- Operationalize SKS-on-Cloud system
- Complete comparative search benchmarks to at least 10 million documents
- Scale to 1 billion, 10 billion, …

Summary

- MIT LL has developed the Structured Knowledge Space system to extract entities and relationships from weakly structured intelligence reporting formats
  - Web services and browser-based user interfaces support discovery and access over the network
- To explore the feasibility and desirability of migrating the full SKS application suite to a cloud-based distributed storage & processing platform, we integrated cloud storage as a data storage sidecar on the existing system
- Early benchmarks indicate that existing system performs adequately up to 3M documents (< 2 sec for simple searches) but timings show an upward trend
  - Too early to predict Cloud-based system performance; however theoretical benchmarks are promising

Acknowledgements

- Gary Condon
- Jason Hepp
- Jeremy Kepner
- Ben Landon
- Bob Piotti
- Chuck Yee
- The LLGrid team
- The SKS-RTRG development team

Contact: Delsey Sherrill, Jonathan Kurz, Craig McNally, and Will Smith

{dsherrill, jonkurz, cmcnally, william.smith}@ll.mit.edu