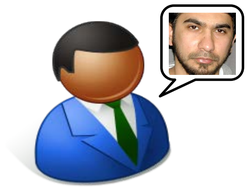
Security and Privacy Assurance Research (SPAR)

Protecting Privacy and Civil Liberties When Querying Sensitive Data

Program Manager: Mr. W. Konrad Vesey; E-mail: william.vesey@iarpa.gov

Undoing the Gordian Knot: when a query is too sensitive to share, and bulk ingestion of the data raises privacy issues

- When a specific question is too sensitive to ask, the solution is often to request everything and extract the desired information in a secure facility
- Data that is ingested in bulk is not kept up to date in real time, requires costly access and retention controls, and risks disclosure of private information due to errors or misbehavior



- Submitting a query to an external party risks disclosure of sensitive or classified selectors

SPAR explored trade-offs among query complexity, security guarantees, processing time and memory requirements

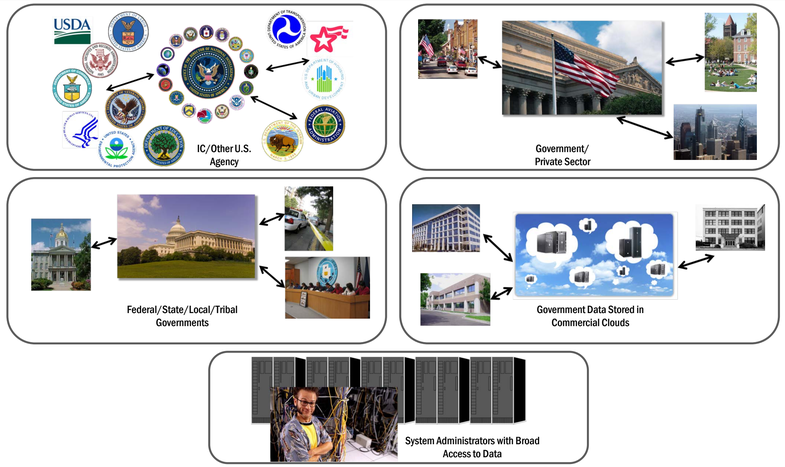
Phase 1	Phase 2
Goal: Query speed comparable to 15x an unprotected query for four query types	Goal: Query speeds comparable to 8x an unprotected query for seven query types

SPAR projects achieved:

- IBM: Query speed comparable to the unprotected MySQL query, but with a storage cost of up to seven times the unencrypted data
- Columbia University: Query speed comparable to seven times as slow as the unprotected MySQL query, but with storage cost of only twice the unencrypted data

Stealth Software Technologies: Query speed that is only practical for queries that return a small number of records (overhead of 600x) but with the strongest security guarantees

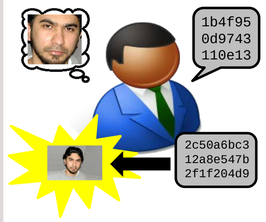
SPAR enables collaboration between non-traditional/occasional partners, and administration without access to content



Information Sharing Applications of SPAR

SPAR protocols give assurance to a data owner that only relevant information is shared

- The data owner does not learn which records were returned
- The data owner verifies that the query complies with an authorization policy—noncompliant queries return no data
- Huge gains in speed are achieved by using an intermediate party to provide storage and computation services
- ***Confidential access to queried data is provided while access to any private information not related to the query is prevented***



SPAR protocols support a practical set of query types and scale to realistic database sizes

Supported Query Types	
Exact Match	Name = “Adam”
Wildcards and Substrings	Name LIKE “Ad*” Name = SUBSTRING(“Adamson”)
Associated Words	WORD_PROXIMITY (“Adam”, “Eve”) <= 100
Word Stems	CONTAINS_STEM (Status, “running”)
Word Search without Predefined Dictionary	CONTAINED_IN (Narrative, “flubber”)
Ranges and Inequalities	Code BETWEEN 100 AND 135 Age >= 18
Boolean Conjunctions of Other Query Terms	Name = “Adam” AND Birthdate <= 10/10/1980 AND NOT Address LIKE “*New York*”
Threshold Conjunctions (must match at least m of n query terms)	M_OF_N (3, 2, Name = “Adam”, Birthdate <= 10/10/1980, NOT Address LIKE “*New York*”)

- SPAR protocols were tested on 10-Terabyte databases with 100 million records

SPAR prototypes are available for evaluation. MIT Lincoln Laboratory can provide application guidance

- MIT LL provided independent test and evaluation of SPAR research prototypes

Questions for future research:

- Beyond database queries, can more general questions, analyses, and functions be efficiently computed on private data without revealing sensitive inputs?
- For a specific application, how can information that is not important to hide be identified and quantified, so that tradeoffs can be explored between efficient performance and information sharing?