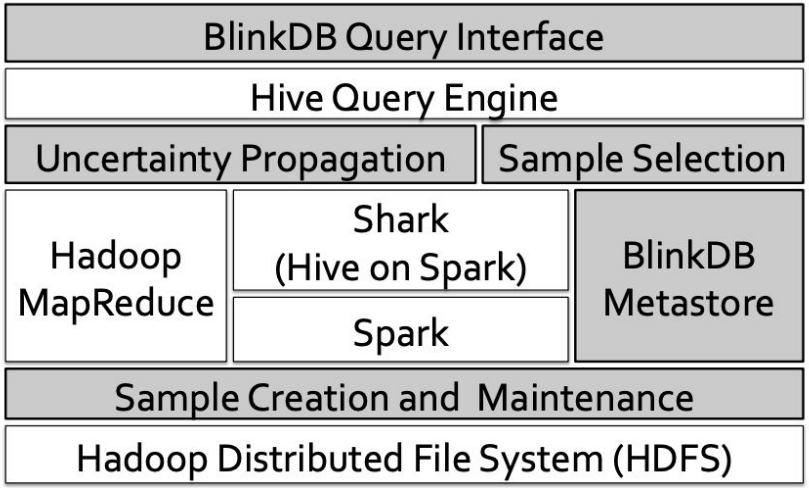
BlinkDB: Queries with Bounded Errors and Bounded Response Times on Very Large Data

Role: Academic Researcher
By: Mayank Sethi

Quick Facts



- It was presented at the EuroSys '13 conf.
- The BlinkDB project was initiated at the **University of California, Berkeley**.
- Addressed the growing need for **interactive, real-time data analytics** on very large datasets.
- Introduced **novel concepts** such as sample-based query processing and user-defined error bounds.

Summary

Introduced techniques like:

- Sample-Based Query Processing
- Error Bounds and User-Defined Tolerance
- Adaptive Query Processing
- Query Scheduling and Resource Management

These techniques allow users to **get query results** within **specified response time constraints** while tolerating a **controlled level of error**. This system significantly enhances the usability of big data for real-time analytics and decision-making.

Scope from the Idea

1. **Dynamic Sampling Strategies:**

   A system that dynamically adjusts the sampling rate based on the query, data distribution, and user-defined error bounds.

2. **Hybrid Query Processing:**

   Develop a system that intelligently switches between approximate and exact processing for different parts of a query or for different types of queries, optimizing accuracy and response times.

Continued:

3. Auto-Tuning of Error using ML:

Develop **machine learning algorithms** that learn from historical query patterns and user preferences to automatically set error bounds for new queries. This would reduce the burden on users and enhance the system's adaptability.

4. Real-Time Data Support:

Extend BlinkDB to handle real-time data sources and streams. Investigate how approximate query processing can be applied to data-in-motion, such as sensor data, social media updates, or financial market feeds.