AsterixDB: A Big Data Management System
“One size fits a Bunch”
https://asterixdb.ics.uci.edu
https://asterixdb.incubator.apache.org

At a Glance
- Open Source (Apache Incubating!) Big Data Management System
- Native support for data ingestion (data feeds)
- Runs on large commodity clusters
- Designed for mass quantities of semi-structured data
- Highly scalable storage and index management
- Native support for rich data types and operations (e.g., spatial & temporal data)
- Native support for similarity queries

AQL (AsterixDB Query Language) + ADM
- ADM is a superset of JSON with a richer set of types (e.g., bags, spatial data, temporal data, text) and optional schemas
- AQL is a powerful query language for semi-structured data, influenced by the best parts of W3C’s XQuery
  ```
  Ex: List the user name and messages sent by those users who joined the Mugshot social network in a certain time frame:
  ```
  ```
  for $user in dataset MugshotUsers
  where $user.user-since >= datet ime(‘2010-07-22T00:00:00’)
  and $user.user-since <= datet ime(‘2012-07-29T23:59:59’)
  return {
  “uname“ : $user.name,
  “messages” : 
  for $message in dataset MugshotMessages
  where $message.author-id = $user.id
  return $message.message
  };
  ```

System Architecture
- Uses the Hyracks data parallel platform as its runtime engine
- Shared-nothing storage
- Built for commodity clusters
- AQL uses Algebricks to optimize queries
- Each Node Controller stores a partitioned portion of the data stored in each index

Data Feeds
- Integrated system for intake of data from live sources (e.g., Twitter)
- Allows for user-defined computation to be performed on ingestion

LSM-Based Storage and Indexing
- Log-Structured Merge technology to support high data ingestion rates
- All primary and secondary index types are LSM-ified (B+Tree, R-Tree, Inverted Indexes, …)

Data-Intensive Computing

Competitive Performance

Sponsored by: