DMF Database Replacement

Arun Ramakrishnan

Technical Lead, Storage Software Engineering

RAIMA Data Manager

- Record based data-store maintained by SGI.
- Maximum of 4 billion stored records with 1 copy and 2 billion with 2 copies of data.
- Supports custom DDL with composite primary keys.
- Snapshots are supported but involve physical copy of table data.
- Indexes are used to locate a record quickly.
- Custom query language implemented using dm*adm tools.
- Data can be imported/exported in text format.
- Dmdbcheck is used to verify database consistency.

RAIMA Data Manager contd ...

- Allows concurrent reader and writer access with support for single writer only.
- No explicit support for foreign keys, table joins etc.
- Multiple record operations can be batched together to improve throughput.
- Operations are journaled in order to facilitate recovery in the event of a crash.
- DMF data maintained across daemon, volume and chunk databases.

Specification for new DB Stack

- Fundamentally distributed design with high level of object count scalability.
- Abstraction of data storage and exchange models.
- Robust toolchain support.
- Language agnostic data abstraction model.
- Ease of porting and database transition.
- Facilitate easy scheduling of Disaster Recovery operations.

NoSQL Stack Evaluation

- Adhere to Eric Brewer's CAP model.
- Inherently distributed in design with user controlling number of replicas.
- Bootstrap, membership management and quorum registration are handled by external stacks.
- Distributed state maintained on a distributed filesystem.
- BigTable and Dynamo based architecture for data modelling.
- Thrift/Avro based IDL with good support for Java and Python.
- Fundamentally looking at Key-Value stores with richer semantics.
- Support concurrent readers and writers to the tables.
- Examples are Cassandra, Hypertable etc.

Relational Stack Evaluation

- Adhere to ACID model with mathematical consistency guarantees.
- Scale to very large volume of data using a few fat nodes.
- Workload can be distributed across sets of nodes using synchronous replication.
- Robust toolchain and language support.
- Cross table dependencies can be modeled and enforced at the database level.

PostgreSQL Features

- Supports very large databases (2 PB+) with 100s of Billions of managed rows.
- Full support for ANSI-SQL:2008 standard.
- Query support for billion row join operations and 100+ GB filters.
- Synchronous replication support for master-standby (hot) style replication.
- Support for multiple programming languages via the PL framework including javascript and python.

Dmaudit data issues

- Dmaudit takes copy of dmf databases and compares them against existing files in filesystem for various attributes.
- Snapshot operation fairly heavy and intrusive since it involves copy of data to alternate location.
- Data has to be sorted, compared and joined multiple times in order to determine errors.
- Dealing with large error counts extremely difficult.
- Current request tracing infrastructure doesn't accommodate DCM msps.
- Input records extracted and compared using FFIO.