A Reconciled Data Warehouse Layer based on CCNx

M. Alexander

Vienna University of Technology

CCNxCon 2013
PARC September 5, 2013

Table of Contents

Problem

Content Centric Networks

Architectural Proposition

Prototype Implementation

Demonstrator Environment

Extensions and Conclusion

Outline

- Very Large DB Data Warehousing Properties
- Content Centric Networking Primitives Congruent with Base Reconciled Data Warehouse Layer Storage Operations
- Basing a Distributed DWH on the CCNx Protocol
- Demonstrate Content Centric MySQL/MariaDB Storage Engine Elements Implementation
- Motivate Broader Content Centric (Super) Distributed Operations

Data Warehousing

- Central Data Repository
  - Variant: Distributed DWH
- Little NoSQL, Mostly Relational
- Data Retrieval: Typical SQL
- 3 Layers
  - Data Sources, Extract, Transform, Load (ETL)
  - Emphasis: Reconciliation Layer
  - DWH, Retrieval Layer
- Distributed Case: Physical Data Mart

Problem Statement

- Large Database Data Warehousing
- Analytic Workloads
- Required Data Locality
  - Bringing the Code to the Data viable Alternative
- Geographically Dispersed Users and Data Provider

How do you Disseminate TB Daily Load Rates to Many Non-Data Center Users Requiring Data Locality?

Content Centric Networks

A Networking Paradigm based on Named Data Entities

Addresses Observed Asymmetry of Internet Traffic Flow Producer-Consumer.

CCNs [13] and the CCNx stack [14] provide primitives that can be applied to the domain of data warehousing (DWH).

- Forwarding
- (IP) Transport Agnostic Communication Interface
- Data Store
- Caching

Content Centric Networking

Properties

- Forwarding via Data Namespace vs. Network Address-based
- Publish/Subscribe on the Network Layers
- Affinity to Delay Tolerant Networks, Content Distribution Networks, Content Addressable Memory and Dataflow
- Maps to Multicast, Anycast
- Data Granularity: Chunks
  - Streaming through Segmentation
- Caching Inherent in CCN Protocol Stack

Content Centric Networking

Properties II

- Current Implementation Utilizes IP Transport
- Interest Packets Trace Content Caching Weights
- Forwarding Tables: Added Pending Interest Table
- Hierarchical (Logical Data) Topologies
  - Source-Sink Cones
- Additional Parameters Relative to TCP/IP
  - Chunk Size, Cache Sizes, Request Rates, Miss Ratios, ...

Explorative Use Cases Survey

- Sequential Scan vs. Index-Based
  - Indexes on all Relations
- Depends on Table Size, Width, Load Rate to Index Build Cost, Results Size of Predicate Evaluation, ...
- Query Plan Cost/Time Experiments Easy to Carry Out using e.g. PostgreSQL EXPLAIN
- Intuitive Ranking of Regions
  - Too Many Dimensions, Tablesize Ranges
  - Distinct Regions with Indexing
    - Local Index + Local Data
    - Local Index + Record ContentObjects
    - Index Retrieved as ContentObject, Records as
  - Distinct Regions without Indexing
    - Convergence on Very Large Analytics DB + Dissemination Case

Use Case

Earth Observation Data Dissemination-Analytics

- Large Vector and Very Large Image Dataset Sizes
  - Fifth Climate Model Intercomparison Project (CMIP) $\backsim 1PB$
  - ESA Earth Observation Data Sets
- High ETL Rates: 100s MB/day
- Suits Hierarchical Dissemination Topology
- Data Locality with Above Parameters Unsolved

Architectural Proposition

Towards a Network-Provided DWH Infrastructure

- Towards DWH Operations on the Network
- Objective of a (Super) Distributed DWH
- Post Grid/Cloud Scientific Data Dissemination/Processing
- Shared-nothing with Explicit Lock Consistency Model
- Global Light-Weight State Holding/Mutex Service
- Flat Node structure
  - But of a Supernode for Query Planning (not in Demonstrator)
  - Index Server (not in Demonstrator)
  - Analyic Segment Servers
- Sharding for Scale-Out

Architectural Proposition II

Towards a Network-Provided DWH Infrastructure

- Proposed Architecture Supports both:
  - Relational
  - Hierarchical
    - Compare Google F1 RDBMS and Spanner Datastore [17]
    - Clusters as Tree Branch Segments for Child Rows

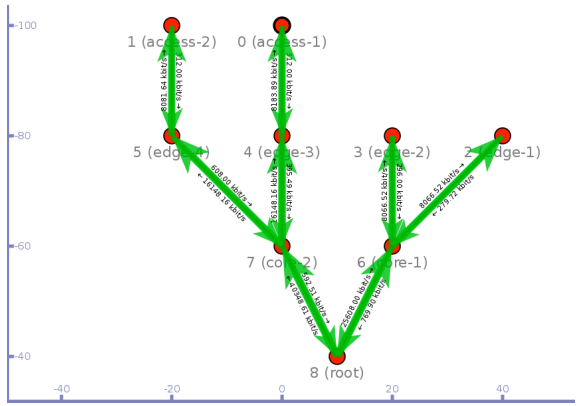
Topology

Topology Options - What is the Optimal Topology for a Table-based CCDB?

- Intuitive Hypothesis as it to be Tree Form (Rooted, Directed Acyclic Graph Shaped)
  - Covering Domain Vector and Image Datasets
- Hard to Test Proposition
- Simulation using Existing Toolchain Enables Parameter Estimation
  - Less Discovery of Optimal Topologies
  - Uses Constructed Skeleton Topology

Simulation

- Regular Throughput, Latency, Hit Ratio Simulation, Tables Repos Upstream/Downstream at Leafs Performed
  - Throughput Shaping as Proxy for Node Utilization
- Possible to Craft Multi-Repository Scenarios based on a-Priori Topology Choice



- Topology Choice Requires Alternate Simulation Approach

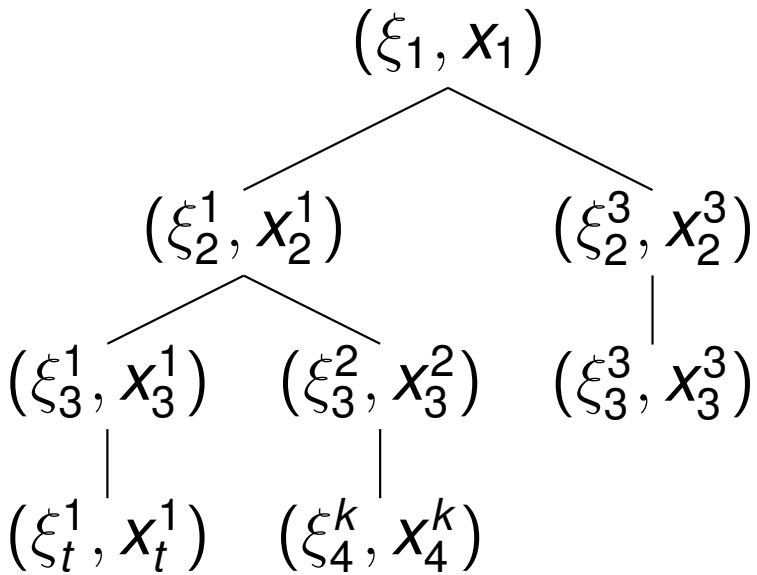
Alternate Optimal Topology Simulation Proposal

Scenario Trees to Track Node Cache Changes over Time

- Proposed Ensemble Method following [9] and [8]
  - Tree as Sample in Time of Stochastic Programming Problem Vector
  - Minimize Cost Along Path
- Procedure
  - Construction of a sample of scenario trees: Monte Carlo and CCN Simulation
  - Solving Sample Trees using Stochastic Programming

Alternate Optimal Topology Simulation (2)

- Procedure contd.
  - Fitting Policy Function using SVM-Machine Learning
  - Ranking and Repeating



Prototype

MariaDB/MySQL

- Storage Engines Extension Options
  - MyISAM with Partitioning
  - CSV: 3 Files per Table, no Indexes
  - (M)Aria with Partitioning
  - Archive: File per Table, Insert-Only, no Indexes
  - InnoDB with Separate Table/Index Files innodb_file_per_table

Prototype

Cross-Layer DB Approaches

- Prior DB Storage Cross-Layer Approaches
  - MySQL Falcon Engine: MySQL over ZFS
    - Management vs. Distributed DB Focus
- DB over DFS
  - MySQL over Distributed {ZFS, GPFS, GluserFS, Lustre, PVFS2 etc.}
  - Locking Issues, Use for Replication,TLOG etc.
- DB over Distributed Block Device
  - MySQL over DRBD et al.

Prototype

Cross-Layer DB Approaches

- Application with DB Protocol over CCN
  - Hadoop with CCNx [16]
- NoSQL over CCN for Large Datasets
  - Distributed Join Cavaet
  - Possibly Implicitly Partitioned, Denormalized
  - or Local Joins

Prototype

Architecture

- MariaDB/MySQL Storage Engine
  - Based on CSV-Engine (file-per-table)
- Engine Modification - New Native Commands
- Global Lock Service
- Modifications for Opening Tables
  - Standard MySQL Behavior: Open/Cache Tablespaces Upon First Operation on Open Database
- Extra Layer of CCDB Global Table Locks
- Explicit Write-back to CCNx from Local FS Store

Prototype

Index-Based Access

- Applicable Regions:
  - Low Change Velocity Tables
  - SELECT Predication Results Sets est. $<<$ 10% Row Numbers
- Initially Examined: Index-Affine Content Object Retrieval based on Local Index
  - Depreciated in Favor of Very Large DB Use Case
  - Where Indexing is Not Efficient vor Very Large Table Sizes, ETL Rates
    - Main Constraint: Index (Re-)Building Time

Prototype Implementation

Design

- MariaDB Database 10.0.4 Fork
- Apache Zookeeper (formerly part of Hadoop) for Global State
- C/C++ CCDB Extensions, Mutex Library

Prototype Implementation

- Ongoing work on Adaption of MySQL/MariaDB CSV Storage Engine
  - Addition of a native MySQL Commands
- Simple Table-Based Granularity
  - Fine Granularity with Multiple Supernodes: Partitioning and Sharding - Towards (Super) Distributed DBs
- Tablespaces
  - Mix of - DB2 Terms - System and Database Managed Space (SMS/DMS) with Filesystem and CCNx

Prototype

Architecture::Hierarchical Namespace

- Hierarchical Demonstrator
- Topology Agnostic but Distribution Tier-Aware
- Maps to CCNx URI Schema
- CCDB Schema
  - ccnx:/ccvldb.org/cryosat2/{altimeter, doppler}/track[n]

Prototype

New MariaDB/MySQL Commands - Partly Implemented

- CCDB_PORT *uri*
- CCDB_OPENTABLE *uri*
- SHOW CCTABLES
- SHOW CCDB_PUBLISH_STATUS
- SHOW CCDB_STATUS uri
- CCDB_PUBLISH *uri*
- CCDB_LOCK_CCTABLE *uri*
- CCDB_UNLOCK_CCTABLE *uri*

Prototype Implementation

Functions Excerpt

```
char* ccdb_mutex_read(std::string uri);
char* ccdb_mutex_check(std::string uri);
std::string ccdb_mutex_lock(std::string uri,
          std::string node, std::string user);
char* ccdb_mutex_release(std::string uri);
char* ccdb_publish_status(std::string uri);
char* ccdb_show_status(std::string uri);
char* ccdb_mutex_createlocknode(std::string uri);
```

Prototype Implementation

In-Memory vector struct Mirroring Zookeper State

```
public:
    ...
    std::vector<std::string> stVector;
    ...
    struct ccdbState{
        std::string uri;
        std::string published;
        std::string node;
        std::string user;
    };
    std::vector<ccdbState> ccdbMutexes;
```

Prototype

Limitations

- No Indexes, Foreign Keys
- Sequential Scan Only
- Single Instance (compare Oracle)
- Single Schema (database.schema.table)
- No DB Owner (Namespace)

Prototype

Walkthrough Excerpts

- SQL SHOW ENGINES;

```
mysql> show engines;
+--------------------+---------+----------------------------------------------------------------+--------------+------+------------+
| Engine             | Support | Comment                                                        | Transactions | XA   | Savepoints |
+--------------------+---------+----------------------------------------------------------------+--------------+------+------------+
| MEMORY             | YES     | Hash based, stored in memory, useful for temporary tables      | NO           | NO   | NO         |
| InnoDB             | DEFAULT | Supports transactions, row-level locking, and foreign keys     | YES          | YES  | YES        |
| PERFORMANCE_SCHEMA | YES     | Performance Schema                                             | NO           | NO   | NO         |
| MRG_MyISAM         | YES     | Collection of identical MyISAM tables                          | NO           | NO   | NO         |
| MyISAM             | YES     | MyISAM storage engine                                          | NO           | NO   | NO         |
| CCDB               | YES     | CCDB CCNx storage engine                                       | NO           | NO   | NO         |
| Aria               | YES     | Crash-safe tables with MyISAM heritage                         | NO           | NO   | NO         |
+--------------------+---------+----------------------------------------------------------------+--------------+------+------------+
7 rows in set (0.00 sec)
```

Prototype

Walkthrough Excerpts

- SQL Check Publish Status (CCNx Repo)
  ccdb_publish_status ccvldb$_cryosat2_altimeter_track1;

```
mysql> ccdb_publish_status ccnx$_ccvldborg_cryosat2_altimeter_track1;
+-----------------------------------------------+----------------+-----------------+-----------+
| CCDB Table                                    | Lock Issued To | Repository Node | Published |
+-----------------------------------------------+----------------+-----------------+-----------+
| ccnx:---ccvldborg---cryosat2---altimeter---track1 | akihiro        | okeanos-1       | 1         |
+-----------------------------------------------+----------------+-----------------+-----------+
1 row in set (0.01 sec)
```

Prototype

Walkthrough Excerpts

- SQL Open CCDB Table ccdb_opentable ccvldb$_cryosat2_altimeter_track1

```
mysql> ccdb_opentable ccvldb$_cryosat2_altimeter_track1;
+-----------------------------------+--------------+----------------------+
| CCDB Table                        | Operation    | Read-back Lock State |
+-----------------------------------+--------------+----------------------+
| ccvldb:_cryosat2_altimeter_track1 | CCNx Open OK | Locked               |
+-----------------------------------+--------------+----------------------+
1 row in set (0.01 sec)
```

Prototype

Walkthrough Excerpts

- SQL Lock/Unlock Table ccdb_lock_cctable ccvldb$_cryosat2_altimeter_track1

```
mysql> ccdb_lock_cctable ccvldb$_cryosat2_altimeter_track1;
+-----------------------------------+-----------+---------------------+
| CCDB Table                        | Operation  | Read-back Lock State |
+-----------------------------------+-----------+---------------------+
| ccvldb:_cryosat2_altimeter_track1 | Table Lock | Locked              |
+-----------------------------------+-----------+---------------------+
1 row in set (0.04 sec)

mysql> ccdb_unlock_cctable ccvldb$_cryosat2_altimeter_track1;
+-----------------------------------+-------------+---------------------+
| CCDB Table                        | Operation    | Read-back Lock State |
+-----------------------------------+-------------+---------------------+
| ccvldb:_cryosat2_altimeter_track1 | Table Unlock | Unlocked            |
+-----------------------------------+-------------+---------------------+
1 row in set (0.01 sec)

mysql>
```

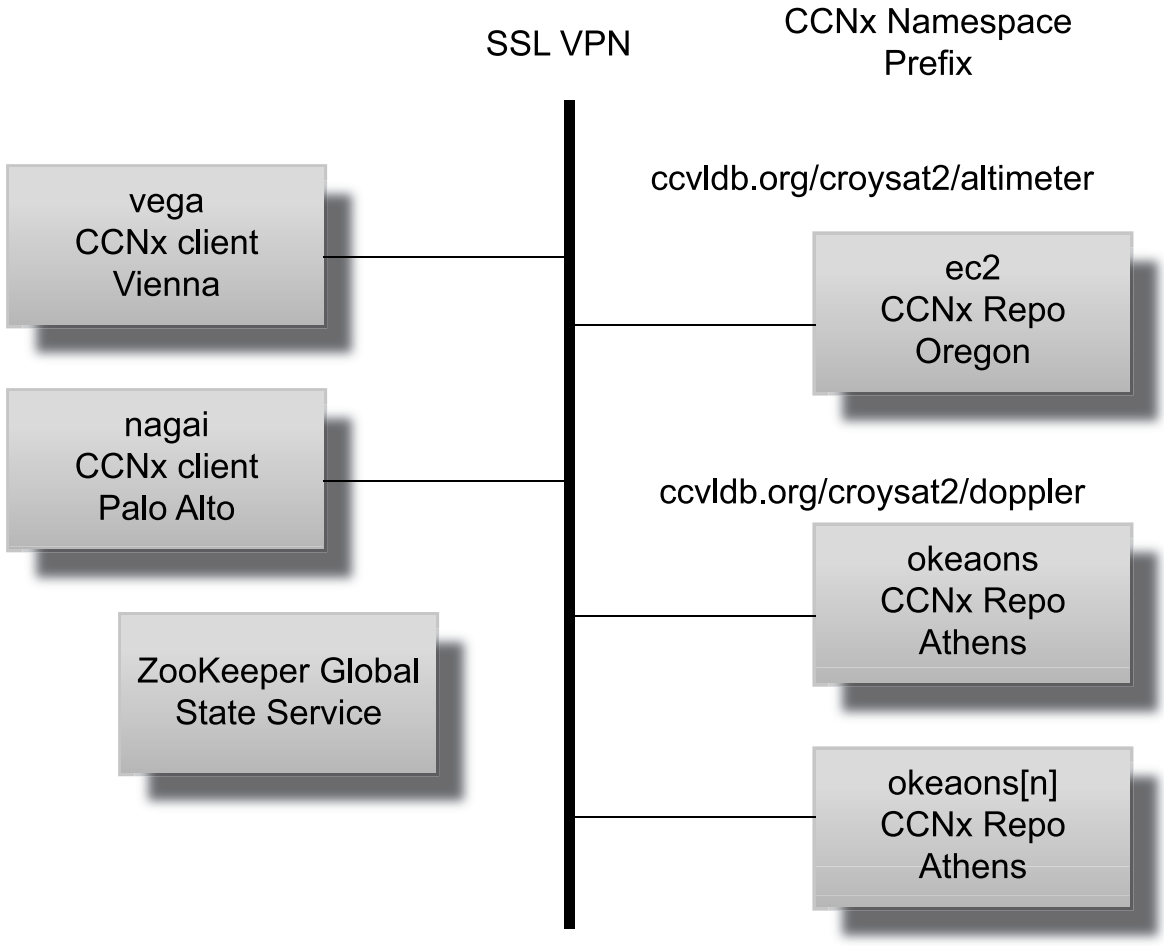
Prototype

Walkthrough Excerpts

- SQL Check Status:
  SHOW CCDB_STATUS

```
mysql> show ccdb_status;
+-------------------------------------------------+---------------+-----------------+-----------+
| CCDB Table                                      | Lock Issued To | Repository Node | Published |
+-------------------------------------------------+---------------+-----------------+-----------+
| ccnx:---ccvldborg---cryosat2---doppler---track1 | dbrootb       | okeanos-2       | 0         |
| ccnx:---ccvldborg---cryosat2---altimeter---track1 | akihiro     | okeanos-1       | 1         |
| ccnx:---ccvldborg---cryosat2---altimeter---track2 | mike1       | ec2             | 1         |
| ccnx:---ccvldborg---cryosat2---altimeter---track3 | akie1       | ec2             | 1         |
| ccnx:---ccvldborg---cryosat2---doppler---track3 | amihiro       | okeanos-3       | 1         |
+-------------------------------------------------+---------------+-----------------+-----------+
5 rows in set (0.02 sec)
```

Demonstrator Environment



Demonstrator

Satellite Remote Sensing Altimeter (Vector) Data

- Croysat-2::Synthetic Interferometric Altimeter (SIRAL)::Low Resolution Mode (LRM)
  - Level 1 Processed
- Data File Split for Relational Join using Matlab Tool (see Acknowledgements)

  C_OFFL_SIR_FDM_1B_20130419T084613_20130419T084845_B001.DBL

Testbed Walkthrough 1

- Open DB
  - use cryosat2_altimeter;
- Check CCDB Status
  - show ccdb_status;
- Open CCDB Table
  - ccdb_opentable ccnx$_ccvldborg_cryosat2_altimeter_lrm;
- Lock CCDB Table
  - ccdb_lock_cctable ccnx$_ccdb_lock_cctable ccnx$_ccvldborg_cryosat2_altimeter_$

Testbed Walkthrough 2

- Selects on Local Table doppler and CCDB Table lrm
  - select * from lrm;
  - select * from doppler;
- Join on them
  - select lrm.a2, doppler.a3 from lrm, doppler where lrm.i=doppler.i;
- Publish Table back to the Network
  - ccdb_publish ccnx$_ccvldborg_cryosat2_altimeter_lrm;
- Check Publish Status
  - ccdb_publish_status

Possible Extensions

Domain Use Case Served Well with Table Granularity

- Lower Granularity
  - Make Clever Use of Segmentation & Seeks?
  - Compare Legacy ISAM Storage Access Method
- Sharding
- Optimistic Concurrency Control

Conclusion

- Proposal of a Large-Data DWH Architecture based on Content Centric Networking
  - Fits OLAP
  - Wide-Area Data Distribution
  - Explicit Read Locality, Write Concurrency Control
- MariaDB/MySQL Integrated Demonstrator

Acknowledgements

- Special thanks to Herbert Findeis
- Salvatore Dinardo (ESA) - Cryosat Matlab Reader Package

References

[]

Alexander Afanasyev.

NS-3 based named data networking (NDN) simulator, March 2013.

Alexander Afanasyev, Ilya Moiseenko, and Lixia Zhang.

ndnSIM: NDN simulator for NS-3.

Technical Report NDN-0005, NDN, October 2012.

Vikas Agrawal, Udayan Nandkeolyar, P.S. Syndararaghavan, and Mesbah Ahmed.

Simulation model and analysis of a data warehouse.

15(3):31–45, 2006.

 **Marcos K. Aguilera, Wojciech Golab, and Mehul A. Shah.**

A practical scalable distributed b-tree.

Proc. VLDB Endow., 1(1):598609, August 2008.

 **Charles Bell.**

***Expert MySQL*.**

Expert's voice in databases. Apress, 2012.

 **Dan Boneh, Amit Sahai, and Brent Waters.**

Functional encryption: a new vision for public-key cryptography.

Commun. ACM, 55(11):5664, November 2012.

Jerome Darmont, Fadila Bentayeb, and Omar Boussaid.

DWEB: a data warehouse engineering benchmark.

CoRR, abs/0705.1453, 2007.

B. Defourny, D. Ernst, and L. Wehenkel.

Scenario trees and policy selection for multistage stochastic programming using machine learning.

ArXiv e-prints, December 2011.

Boris Defourny, Ernst Damien, and Louis Wehenkel.

forthcoming: Scenario trees and policy selection for multistage stochastic programming using machine learning.

INFORMS Journal on Computing, 2012.

Mark Gritter and David R. Cheriton.

An architecture for content routing support in the internet.

In *Proceedings of the 3rd conference on USENIX Symposium on Internet Technologies and Systems - Volume 3*, USITS'01, page 44, Berkeley, CA, USA, 2001. USENIX Association.

Patrick Hunt, Mahadev Konar, Flavio P. Junqueira, and Benjamin Reed.

ZooKeeper: wait-free coordination for internet-scale systems.

In *Proceedings of the 2010 USENIX conference on USENIX annual technical conference*, USENIXATC'10, page 1111, Berkeley, CA, USA, 2010. USENIX Association.

Van Jacobson, Marc Mosko, Diana Smetters, and J.J. Garcia-Luna-Aceves.

Content-centric networking.

Whitepaper, Palo Alto Research Center, January 2007.

Van Jacobson, Diana K. Smetters, James D. Thornton, Michael F. Plass, Nicholas H. Briggs, and Rebecca L. Braynard.

Networking named content.

In *Proceedings of the 5th international conference on Emerging networking experiments and technologies*, CoNEXT '09, page 112, New York, NY, USA, 2009. ACM.

{**Palo Alto Research Center**}.

CCNx reference implementation.

Technical report, PARC, Palo Alto, 2013.

Diego Perino and Matteo Varvello.

A reality check for content centric networking.

In *Proceedings of the ACM SIGCOMM workshop on Information-centric networking*, ICN '11, page 4449, New York, NY, USA, 2011. ACM.

John Sherwood.

An implementation of content-centric networking socket for use with hadoop.

Technical report, Stetson University, 2011.

Jeff Shute, Mircea Oancea, Stephan Ellner, Ben Handy, Eric Rollins, Bart Samwel, Radek Vingralek, Chad Whipkey, Xin Chen, Beat Jegerlehner, Kyle Littleeld, and Phoenix Tong.

F1 - the fault-tolerant distributed RDBMS supporting google's ad business.

In *SIGMOD*, 2012.

Talk given at SIGMOD 2012.