Cassandra Essentials Tutorial Series

An Overview of Apache Cassandra

Agenda

- What is Cassandra?
- History
- Architecture
- Key Features and Benefits
- Who's using Cassandra?
- Where to get Cassandra

Definition of Cassandra

Apache Cassandra™ is a free

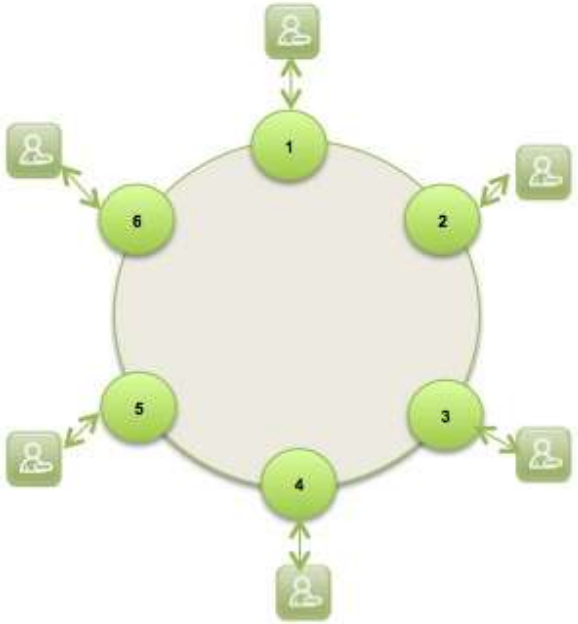
Distributed…
High performance…
Extremely scalable…
Fault tolerant (i.e. no single point of failure)…

post-relational database solution. Cassandra can serve as both real-time datastore (the "system of record") for online/transactional applications, and as a read-intensive database for business intelligence systems.

The History of Cassandra

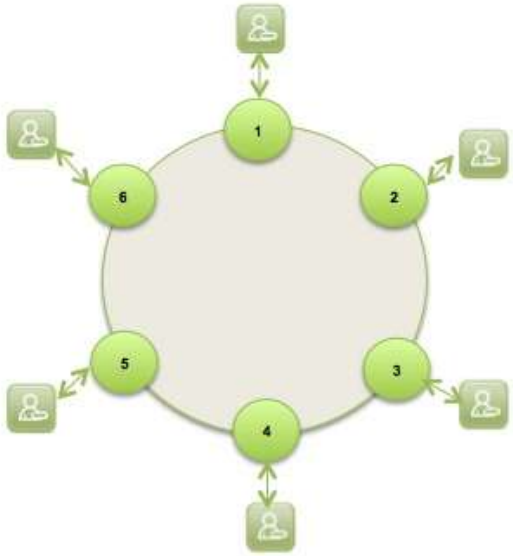
Architecture Overview

- Cassandra was designed with the understanding that system/hardware failures can and do occur
- Peer-to-peer, distributed system
- All nodes the same
- Data partitioned among all nodes in the cluster
- Custom data replication to ensure fault tolerance
- Read/Write-anywhere design



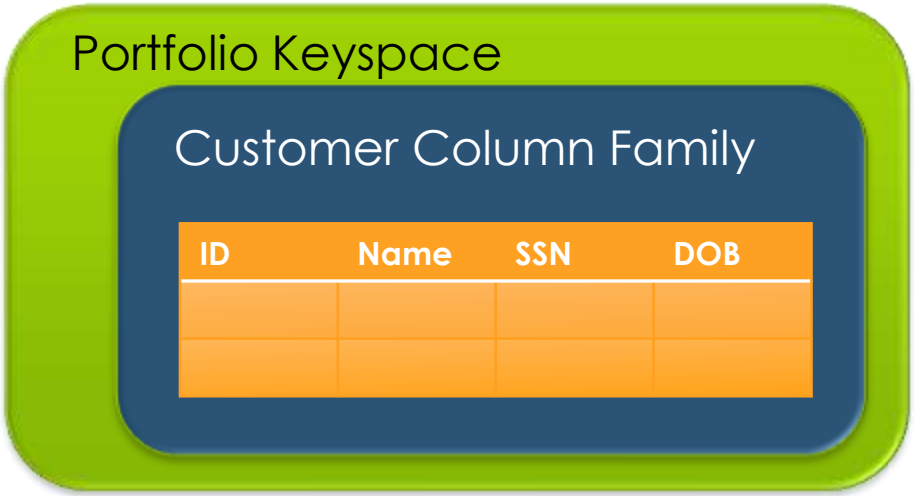
Architecture Overview

- Each node communicates with each other through the Gossip protocol, which exchanges information across the cluster every second
- A commit log is used on each node to capture write activity. Data durability is assured
- Data also written to an in-memory structure (memtable) and then to disk once the memory structure is full (an SStable)



Architecture Overview

- The schema used in Cassandra is mirrored after Google Bigtable. It is a row-oriented, column structure
- A keyspace is akin to a database in the RDBMS world
- A column family is similar to an RDBMS table but is more flexible/dynamic
- A row in a column family is indexed by its key. Other columns may be indexed as well

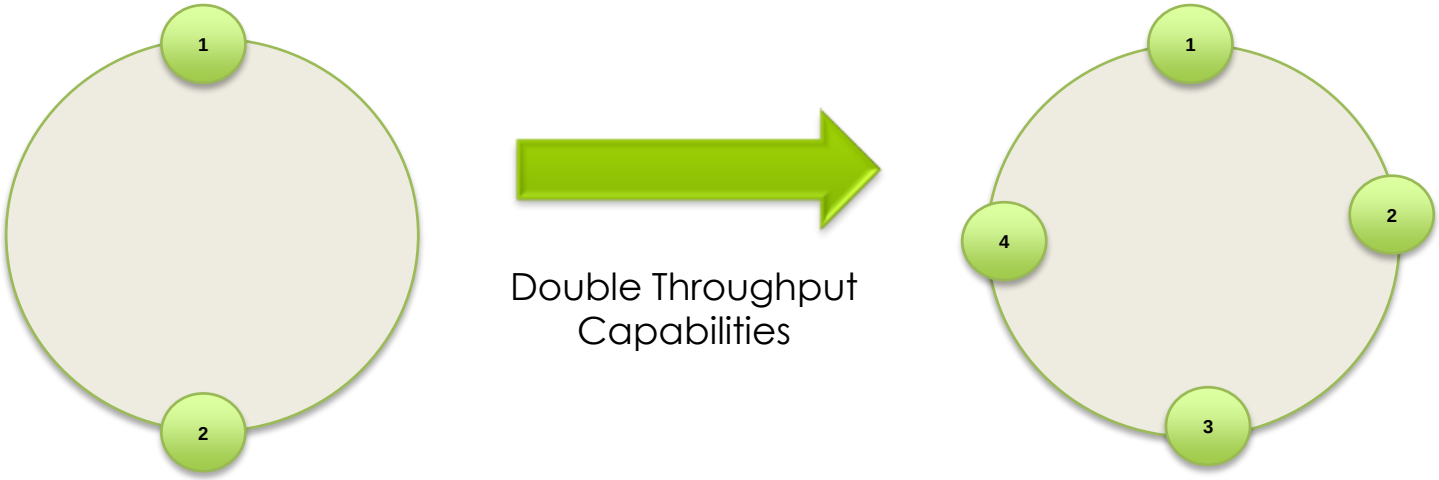


Why Cassandra?

- Gigabyte to Petabyte scalability
- Linear performance gains through adding nodes
- No single point of failure
- Easy replication / data distribution
- Multi-data center and Cloud capable
- No need for separate caching layer
- Tunable data consistency
- Flexible schema design
- Data Compression
- CQL language (like SQL)
- Support for key languages and platforms
- No need for special hardware or software

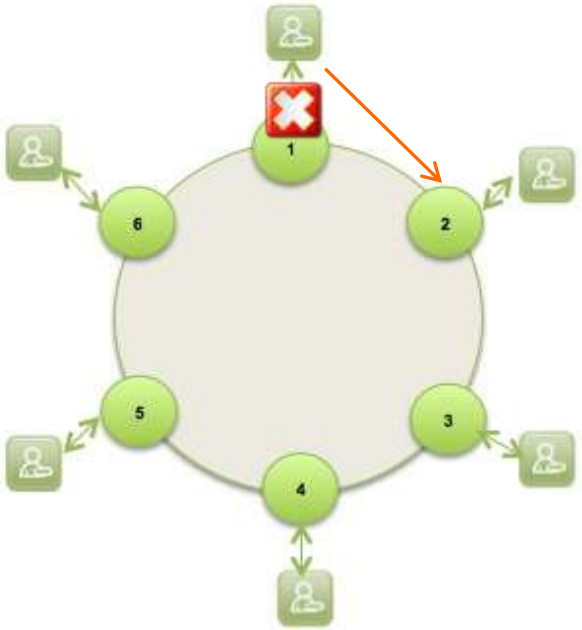
Big Data Scalability

- Capable of comfortably scaling to petabytes
- New nodes = Linear performance increases
- Add new nodes online



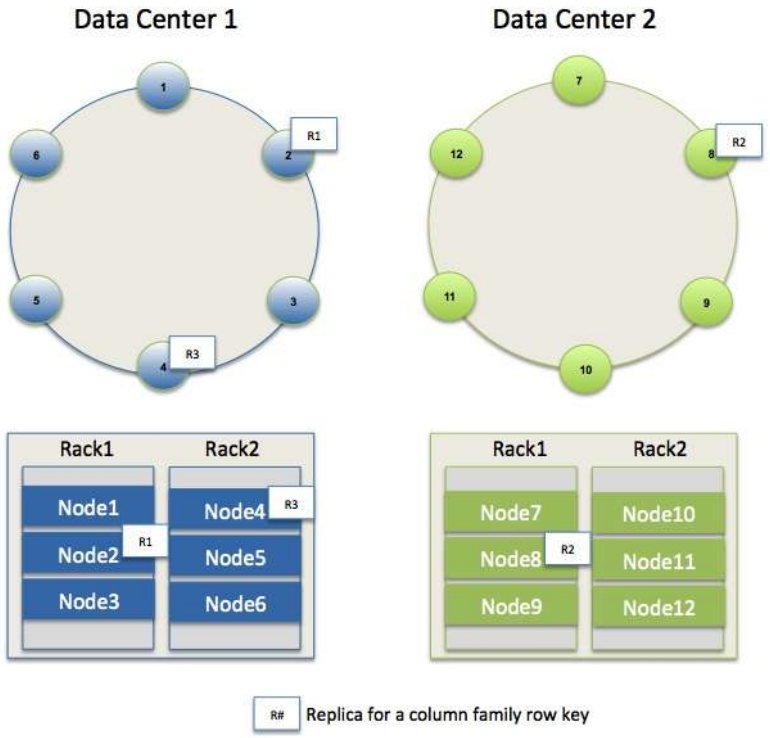
No Single Point of Failure

- All nodes the same
- Customized replication affords tunable data redundancy
- Read/write from any node
- Can replicate data among different physical data center racks



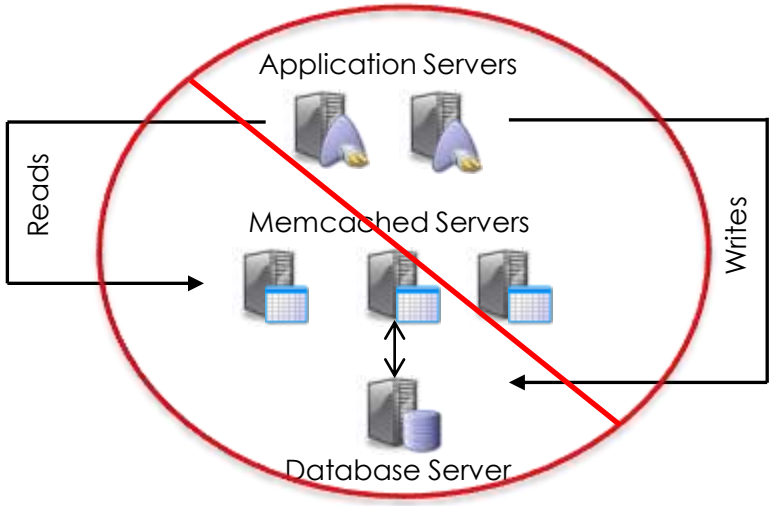
Easy Replication / Data Distribution

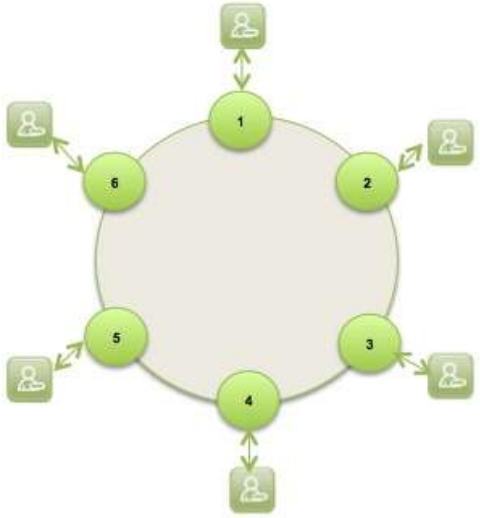
- Transparently handled by Cassandra
- Multi-data center capable
- Exploits all the benefits of Cloud computing
- Able to do hybrid Cloud/On-premise setup



No Need for Caching Software

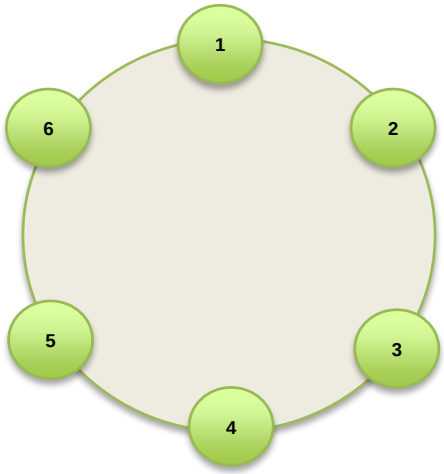
- Peer-to-peer architecture removes need for special caching layer and the programming that goes with it
- The database cluster uses the memory from all participating nodes to cache the data assigned to each node
- No irregularities between a memory cache and database are encountered





Tunable Data Consistency

- Choose between strong and eventual consistency (All to any node responding) depending on the need
- Can be done on a per-operation basis, and for both reads and writes
- Handles Multi-data center operations



Writes

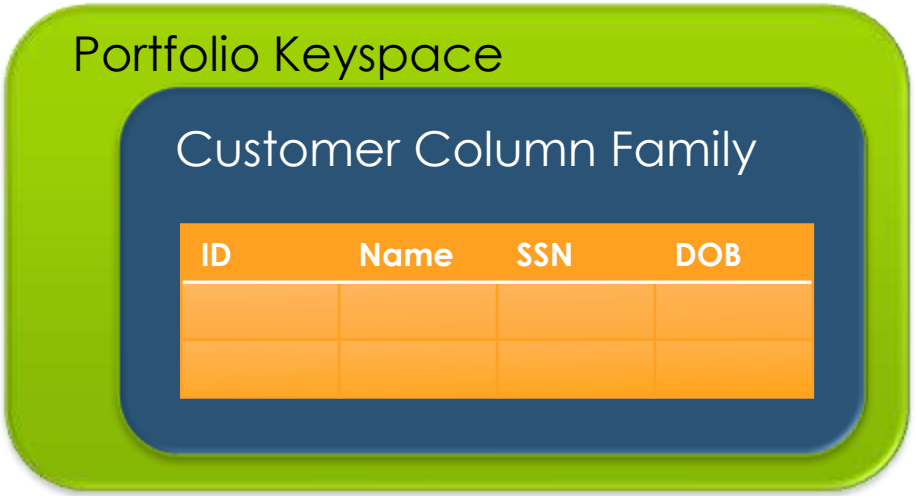
- Any
- One
- Quorum
- Local_Quorum
- Each_Quorum
- All

Reads

- One
- Quorum
- Local_Quorum
- Each_Quorum
- All

Flexible Schema

- Dynamic schema design allows for much more flexible data storage than rigid RDBMS
- Handles structured, semi-structured, and unstructured data. Counters also supported
- No offline/downtime for schema changes
- Supports primary and secondary indexes

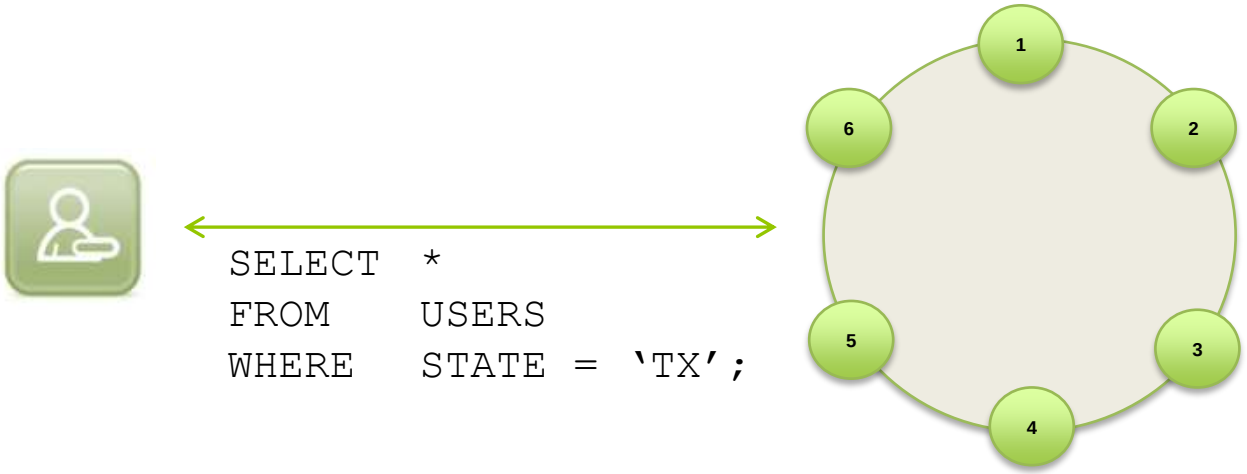


Data Compression

- Uses Google's Snappy data compression algorithm
- Compresses data on a per column family level
- Internal tests at DataStax show up to 80%+ compression of raw data
- No performance penalty (and some increases in overall performance due to less physical I/O)!

CQL Language

- Very similar to RDBMS SQL syntax
- Create objects via DDL (e.g. CREATE…)
- Core DML commands supported: INSERT, UPDATE, DELETE
- Query data with SELECT



Who's Using Cassandra?

http://www.datastax.com/cassandrausers#all

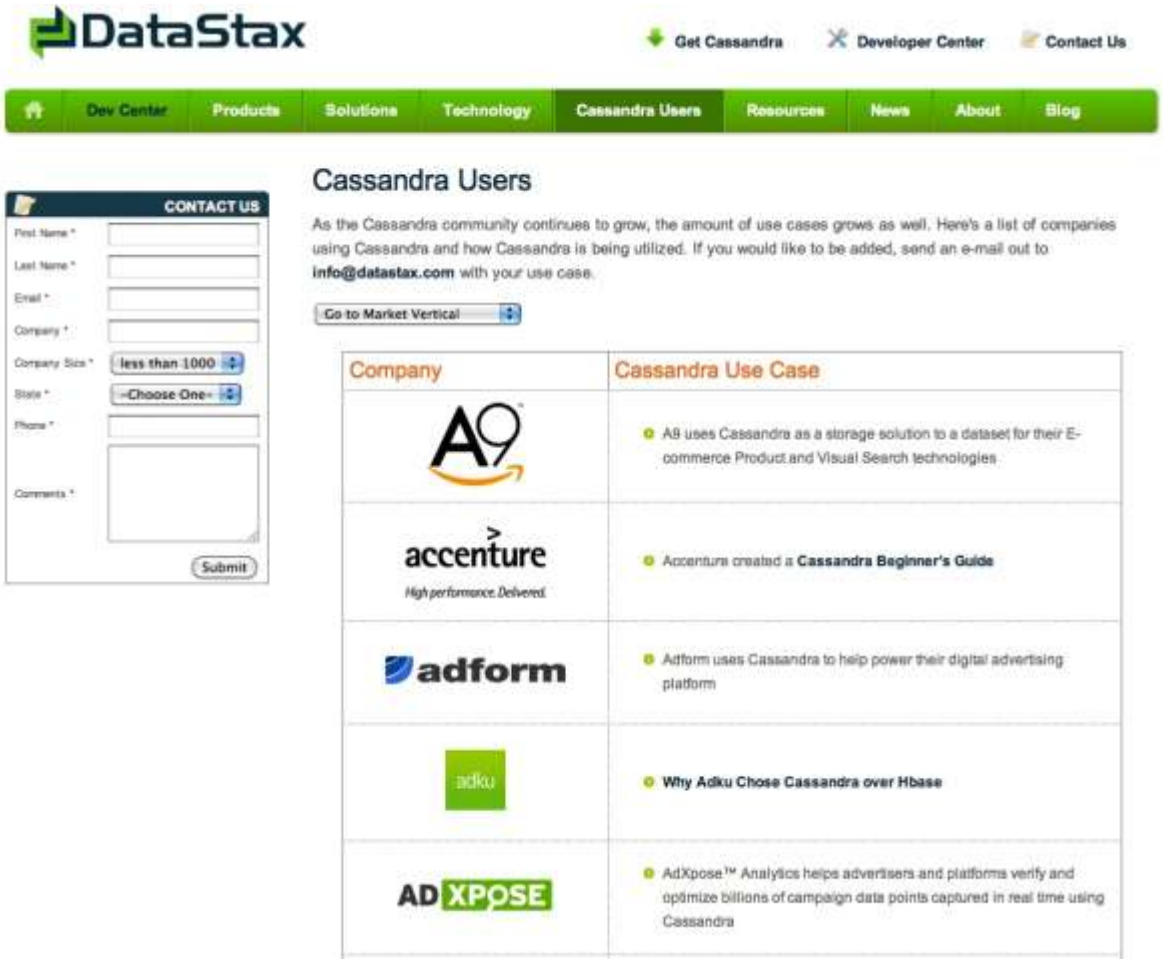
NETFLIX

twitter

cisco

hp invent

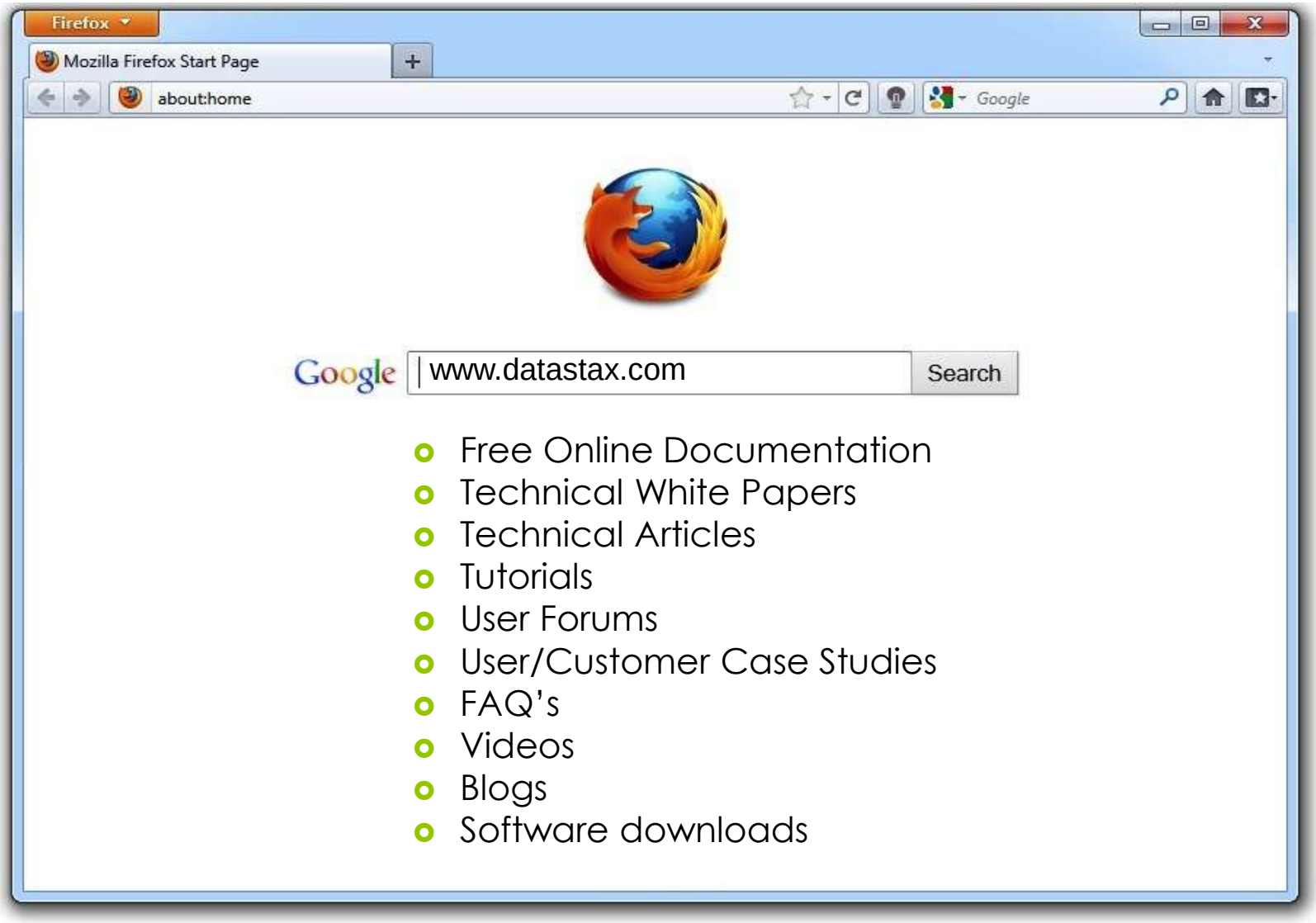
UBS



Where to get Cassandra?

- Go to www.datastax.com
- DataStax makes free smart start installers available for Cassandra that include:
  - The most up-to-date Cassandra version that is production quality
  - A version of DataStax OpsCenter, which is a visual, browser-based management tool for managing and monitoring Cassandra
  - Drivers and connectors for popular development languages
  - Same database and application
  - Automatic configuration assistance for ensuring optimal performance and setup for either stand-alone or cluster implementations
  - Getting Started Guide

Where Can I Learn More?



- Free Online Documentation
- Technical White Papers
- Technical Articles
- Tutorials
- User Forums
- User/Customer Case Studies
- FAQ's
- Videos
- Blogs
- Software downloads

Cassandra Essentials Tutorial Series

An Overview of Apache Cassandra

Thanks…!