Making Hadoop Real Time with **Scala & GridGain**

Nikita Ivanov, Founder & CEO
September 2012

GridGain Systems
www.gridgain.com

1065 East Hillsdale Blvd, Suite 230
Foster City, CA 94404

#gridgain

Table Of Contents:

- **30%**
  - Why Real Time Hadoop?
  - GridGain Overview
  - In-Memory Computing
  - Compute & Data Grids
- **70%**
  - Live Coding
  - Real Time & Streaming Word Count

Real Time Hadoop:

GridGain Overview:

- **Scalable In-Memory Data Platform**
  In-Memory Compute Grid + In-Memory Data Grid
  Real Time & Streaming MapReduce, CEP
- **Three Editions:**
  - "Compute Grid" Edition
    Targeted at HPC market
  - "Data Grid" Edition
    Targeted at Transactional Data Caching market
  - "Big Data" Edition
    Targeted at Real Time Big Data market
- Language support:
  Server: Java, Scala, Groovy,
  Clients: .NET, PHP, REST, C++
- Mobile platforms support:
  iOS/ObjectiveC, Android clients

- **Full ACID**
  Fully distributed ACID transactions
- **Simplicity** and **Productivity**
  Dramatically reduces cost of application development
  Demonstrably faster time-to-market

Example:

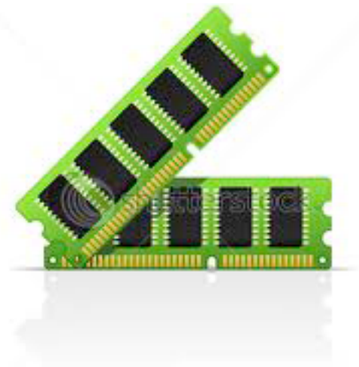
Full source code in Scala of **world's shortest** real time MapReduce app built with GridGain. Works on **one** or **thousands** of computers with **no code changes** required.

```
1  import org.gridgain.scalar.scalar
2  import scalar._
3
4  object MapReduce extends App {
5      scalar {
6          println("Count is: " +
7              grid$.spreadReduce(for (w <- args(0).split(" "))
8                  yield () => w.length)(_.sum)
9          )
10     }
11 }
```

In-Memory Computing: Why Now?

“In-memory will have an industry impact comparable to web and cloud. RAM is a new disk, and disk is a new tape.”

Gartner



Technology & Cost:

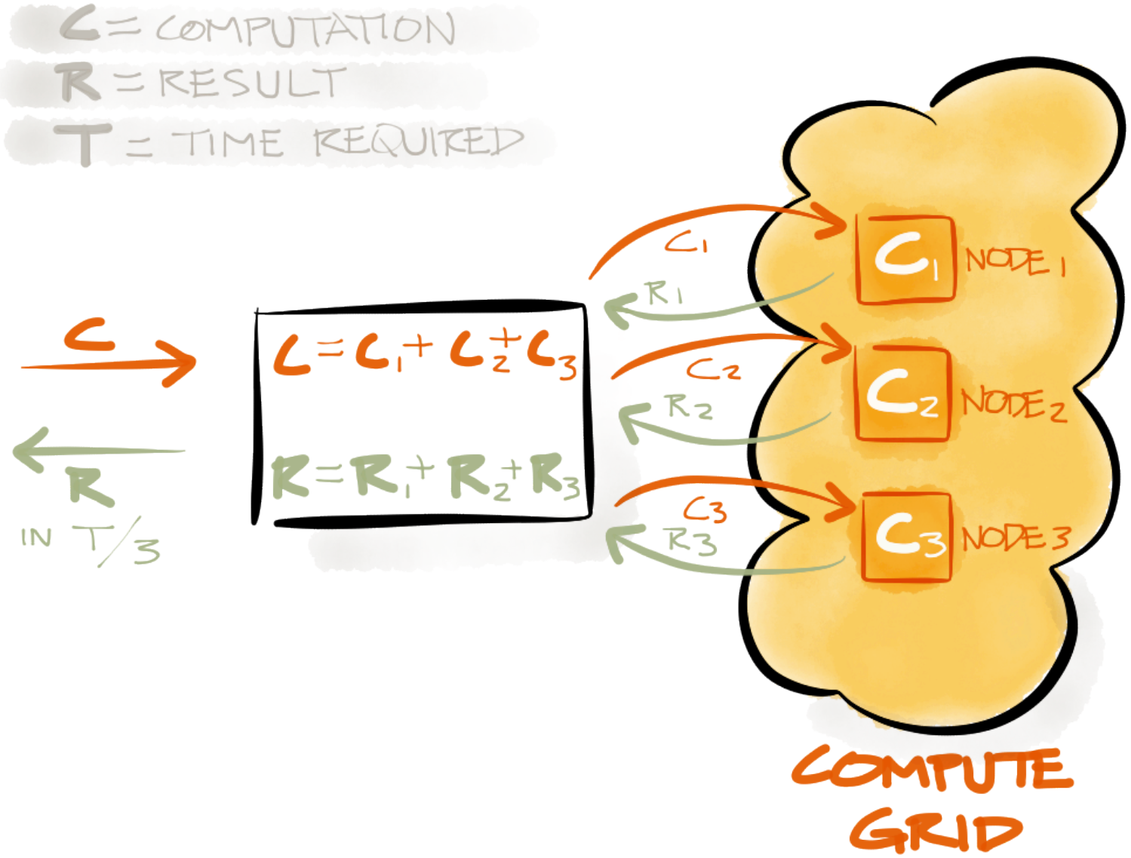
- 64-bit CPU can address **16 exabytes**
  Entire active data set on the planet is addressable by just 1 CPU.
- Disk up to **10^5 times slower** than DRAM
  SSD drives are up to 10^3 times slower
- Super effective **in-memory parallelization**
  Enabled by modern multicore CPUs
- DRAM prices drop **30% every 18 months**
  1TB RAM & 48 cores cluster ~ $40K (< $20K in 3 years)

Performance & Scalability Matters:

- **Citi:** 100ms == $1M loss
  Forex trading
- **Google:** 500ms == 20% traffic drop
  Dropping 20% of revenue
- **SAP** sees +206% in profit in Q112
  For in-memory SAP HANA products
- **Software AG** sees 3x revenue in 2012
  For in-memory Terracota products

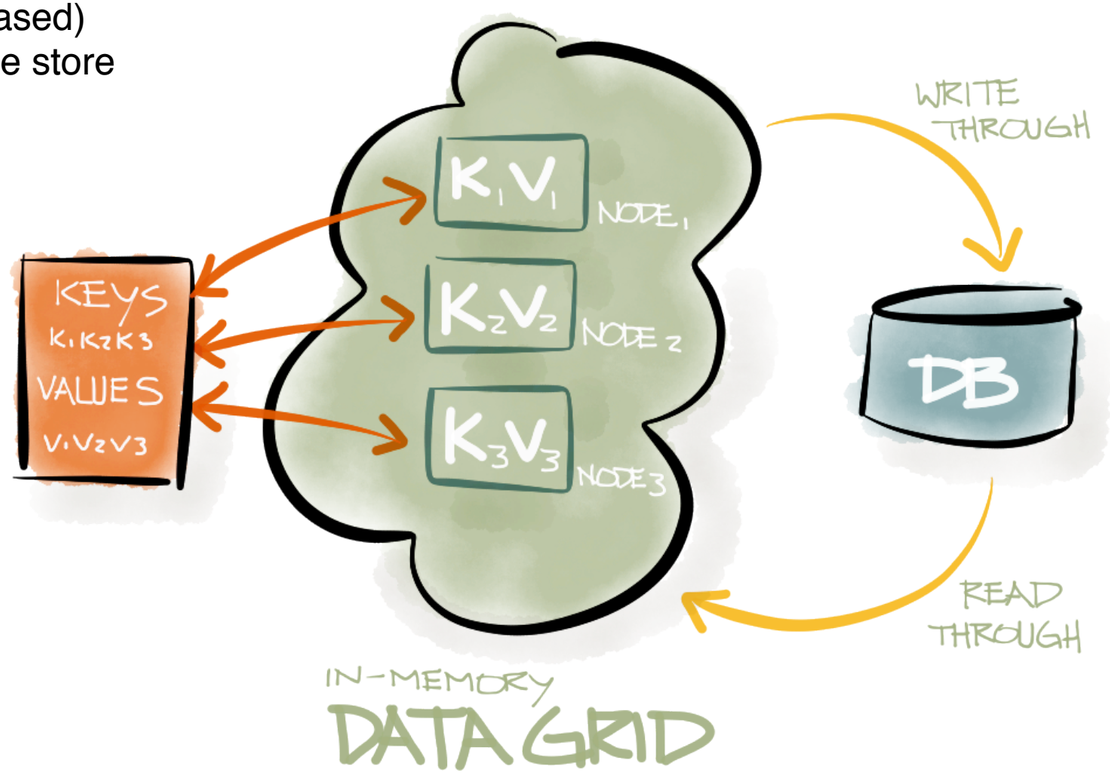
GridGain: In-Memory Compute Grid

- Direct API for split and aggregation
- Pluggable failover, topology and collision resolution
- Distributed task session
- Distributed continuations & recursive split
- Support for Streaming MapReduce
- Support for Complex Event Processing (CEP)
- Node-local cache
- AOP-based, OOP/FP-based execution modes
- Direct closure distribution in Java, Scala and Groovy
- Cron-based scheduling
- Direct redundant mapping support
- Zero deployment with P2P class loading
- Partial asynchronous reduction
- Direct support for weighted and adaptive mapping
- State checkpoints for long running tasks
- Early and late load balancing
- Affinity routing with data grid



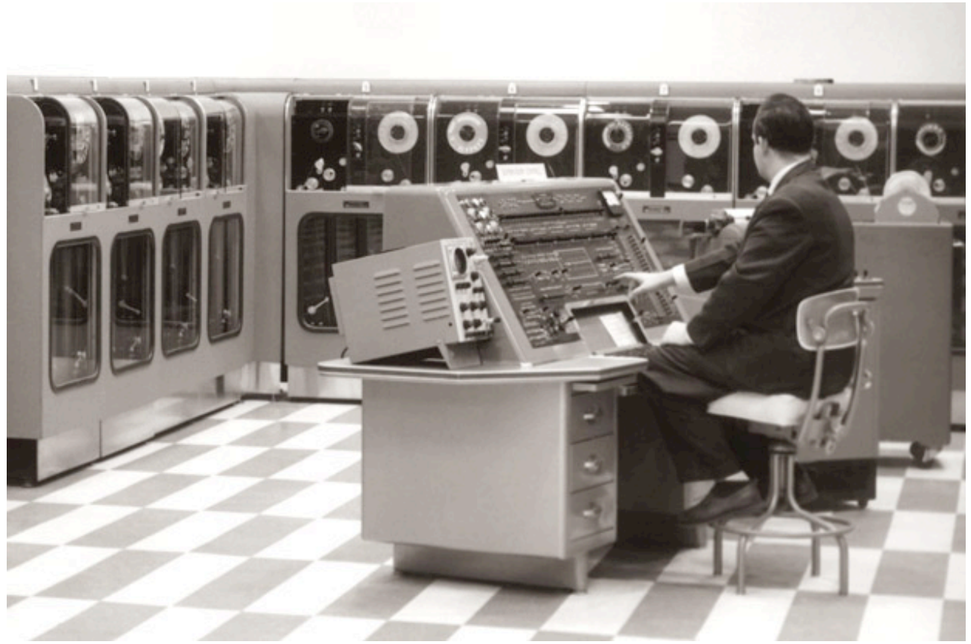
GridGain: In-Memory Data Grid

- Zero deployment for data
- Local, full replicable and partitioned cache types
- Pluggable expiration policies (LRU, LIRS, random, time-based)
- Read-through and write-through logic with pluggable cache store
- Synchronous and asynchronous cache operations
- MVCC-based concurrency
- Pluggable data overflow storage via new swap space SPI
- PESSIMISTIC, OPTIMISTIC transactions
- Standard isolation levels, JTA/JCA integration
- Master/Master data replication/invalidation
- Write-behind cache store support
- Concurrent and transactional data preloading
- Delayed preloading support
- Affinity routing with compute grid
- Partitioned cache with active replicas
- Structured and unstructured data
- Datacenter replication
- JDBC driver for in-memory object data store
- Off-heap memory support
- Pluggable indexing via Indexing SPI
- Tiered storage with on-heap, off-heap, swap space, SQL, and Hadoop
- Distributed in-memory query capability
- SQL, H2, Lucene, predicate-based affinity co-located queries



Live Coding: GridGain + Scala

- **100% Live Coding:**
  - Nothing pre-built
  - Every line & character
  - Everything from the start



GridGain

IN-MEMORY BIG DATA

Thank You!

#gridgain