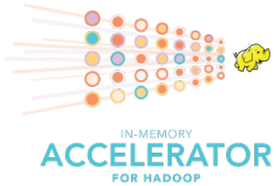
In-Memory Accelerator for Hadoop™



GridGain
IN-MEMORY COMPUTING

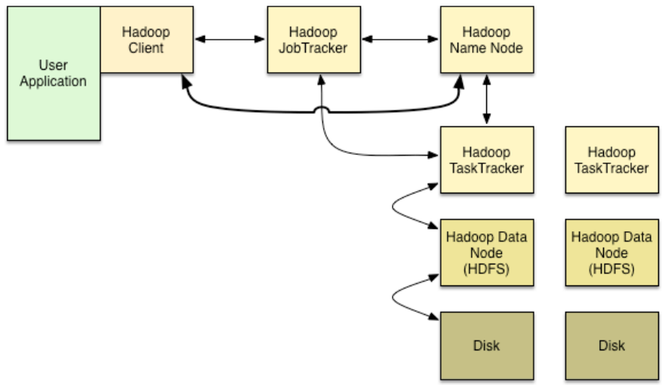
www.gridgain.com

#gridgain

Hadoop: Pros and Cons

What is Hadoop?

- Hadoop is a batch system
- HDFS - Hadoop Distributed File System
- Data must be ETL-ed into HDFS
- Parallel processing over data in HDFS
- Hive, Pig, HBase, Mahout...
- Most popular data warehouse



Pros:

- Scales very well
- Fault tolerant and resilient
- Very active and rich eco-system
- Process TBs/PBs in parallel fashion

Cons:

- Batch oriented - **real time not possible**
- Complex deployment
- Significant execution overhead
- HDFS is IO and network bound

In-Memory Accelerator For Hadoop: Overview

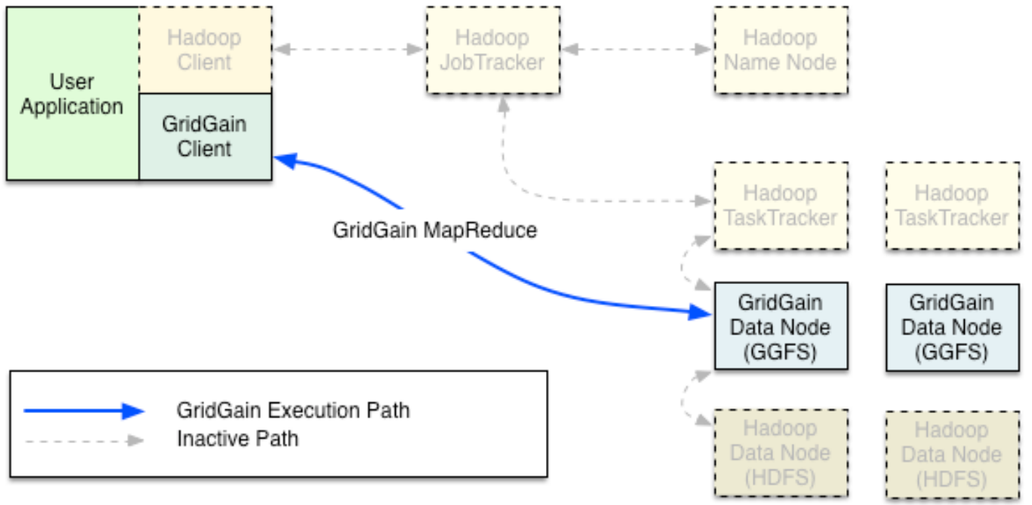
Up To 100x Faster:

1. In-Memory File System

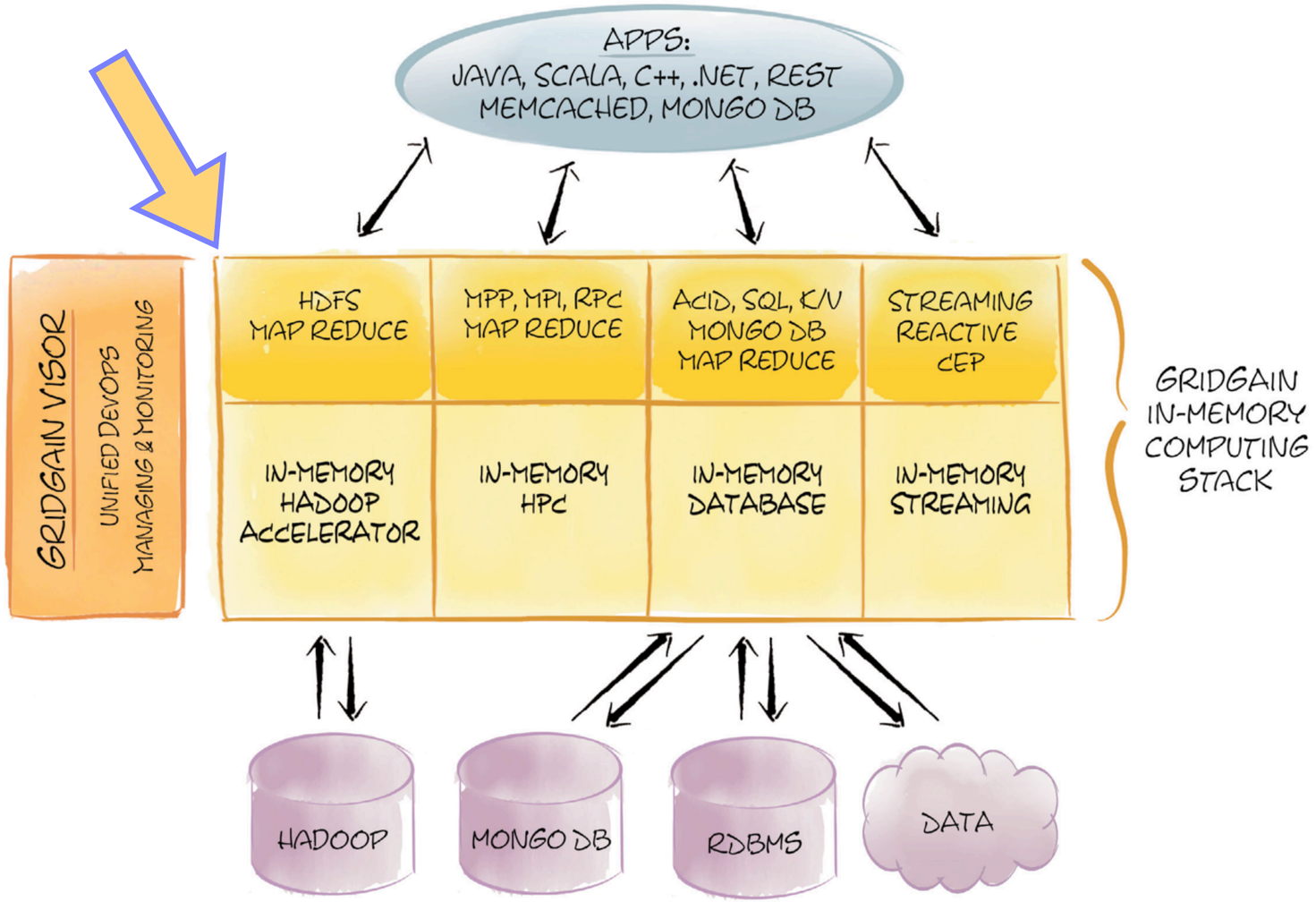
100% compatible with HDFS
Boost HDFS performance by removing IO overhead
Dual-mode: standalone or caching
Blend into Hadoop ecosystem

2. In-Memory MapReduce

Eliminate Hadoop MapReduce overhead
Allow for embedded execution
Record-based



GridGain: In-Memory Computing Platform



In-Memory Accelerator For Hadoop: Details

PnP Integration

Minimal or zero code change

Any Hadoop distro

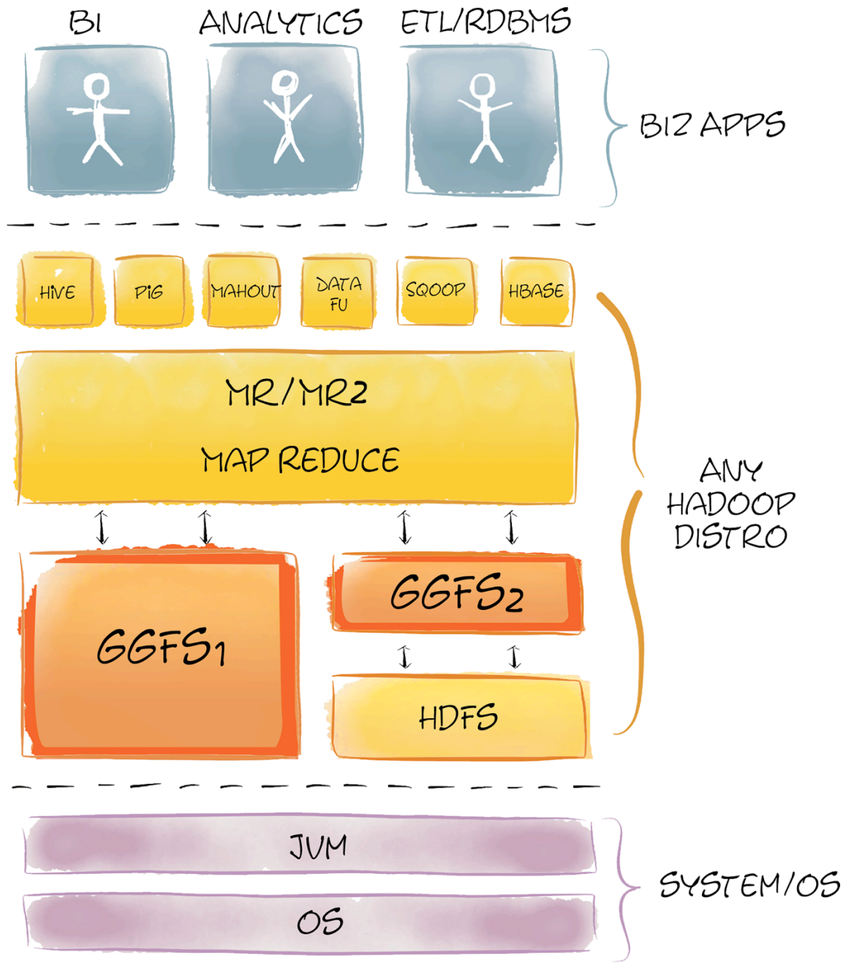
Hadoop v1 and v2

In-Memory File System

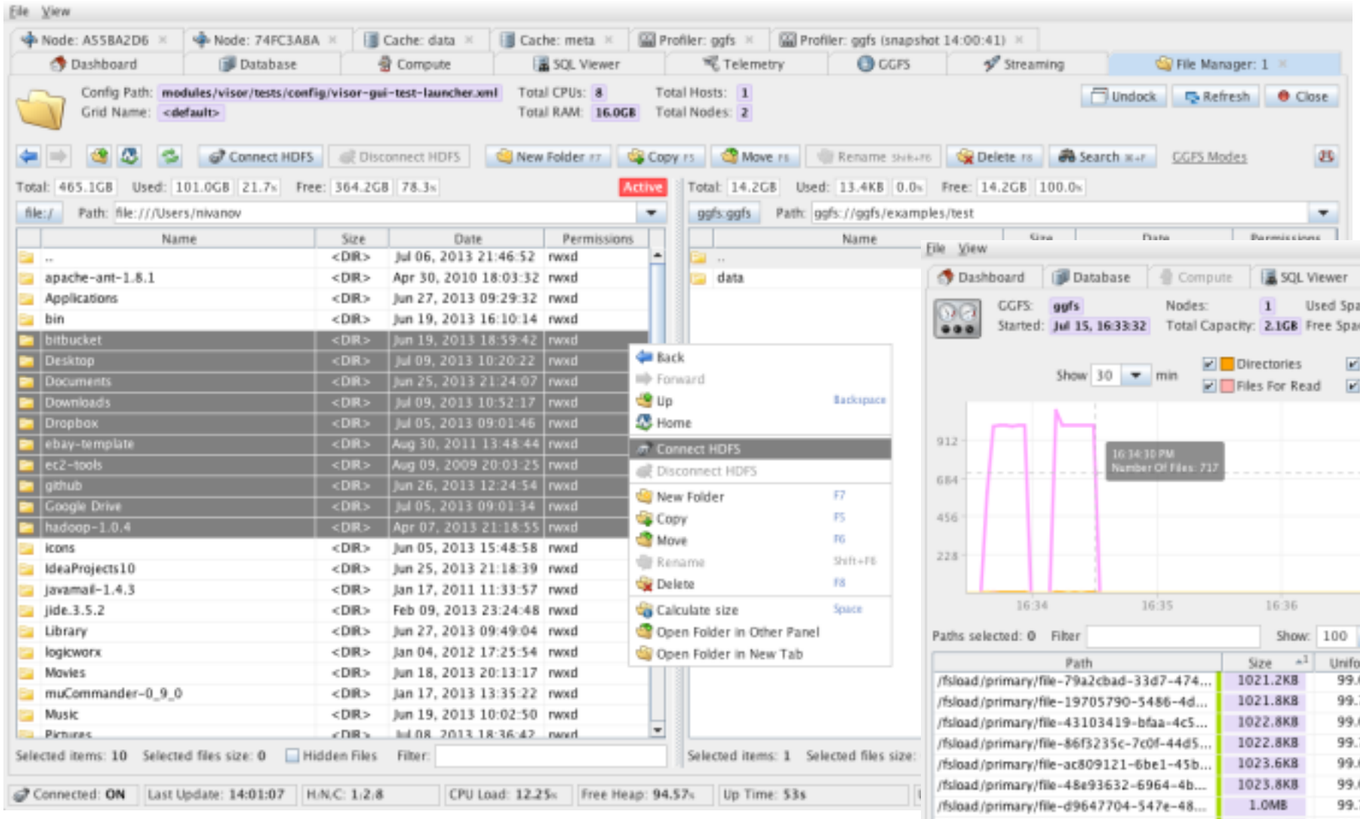
100% compatible with HDFS
Dual-mode: no ETL needed, read/write-through
Block-level caching & smart eviction
Automatic pre-fetching
Background fragmentizer
On-heap and off-heap memory utilization

In-Memory MapReduce

In-process co-located computations - access GGFS in-process
Eliminate unnecessary IPC
Eliminate long task startup time
Eliminate mandatory sorting and re-shuffling on reduction

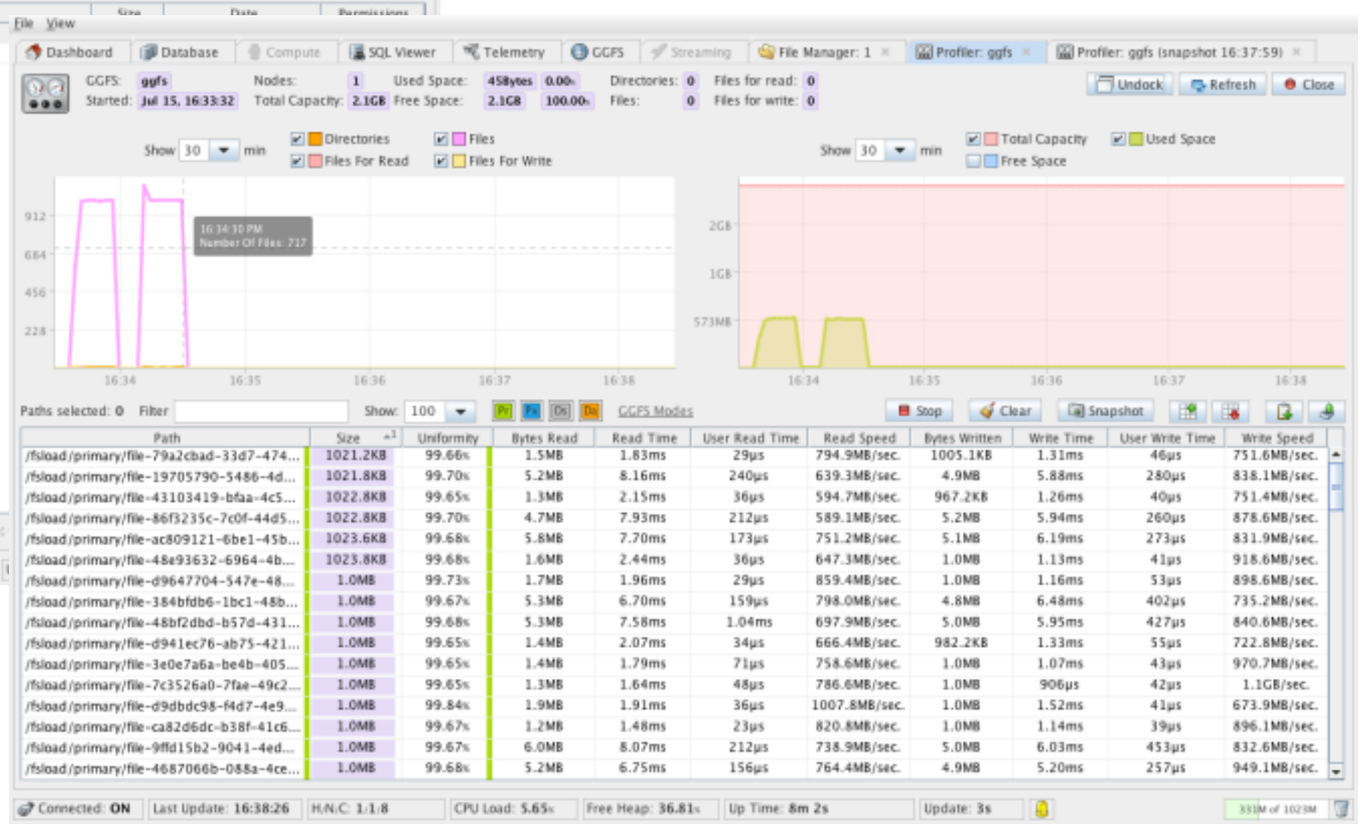


GridGain Visor: Unified DevOps



File Manager

HDFS Profiler



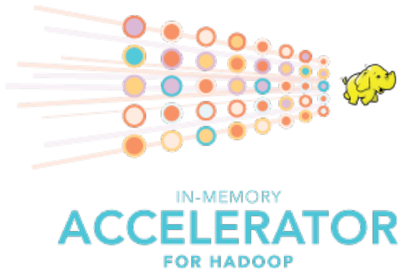
Benchmarks: GGFS vs HDFS

BENCHMARK	GGFS, MS.	HDFS, MS.	BOOST, %
File Scan	27	667	2470%
File Create	96	961	1001%
File Random Access	413	2931	710%
File Delete	185	1234	667%

10 nodes cluster of Dell R610

- Each has dual 8-core CPU
- Ubuntu 12.4, Java 7
- 10 GBE network
- Stock Apache Hadoop 2.x

Comparison: Hadoop Accelerator vs. Spark



- No ETL required
  - Automatic HDFS read-through and write-through
  - Data is loaded on demand
- Per-block file caching
  - Only hot data blocks are in memory
- Strong management capabilities
  - GridGain Visor - Unified DevOps

- Requires data ETL-ed into Spark
  - Changes to data do not get propagated to HDFS
  - Explicit ETL step consumes time
- Needs to have full file loaded
  - If does not fit - gets offloaded to disk
- No management capabilities

Customer Use Case: Task & Challenge

Task:

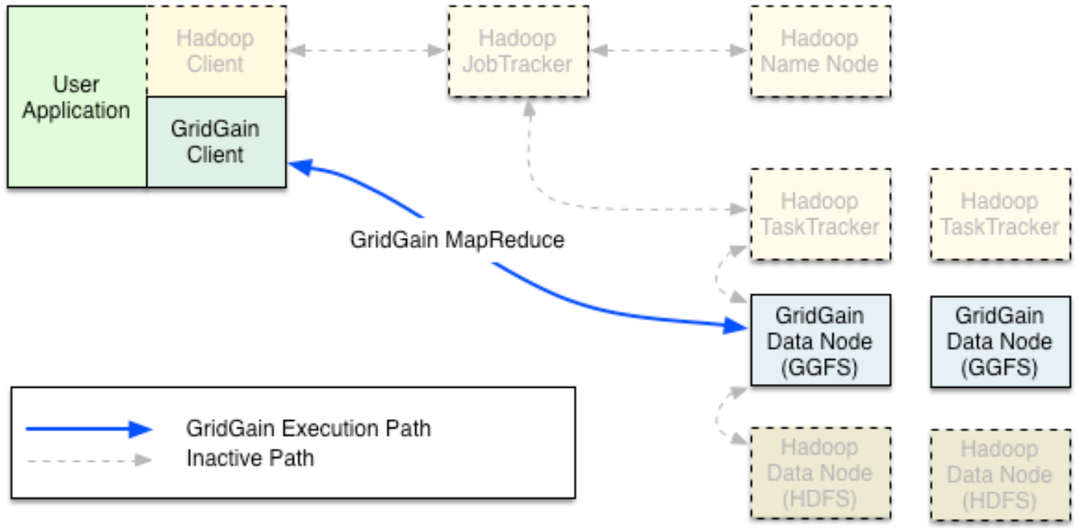
- Real time search with MapReduce
- Dataset size is 5TB
- Writes 80%, reads 20%
- Perceptual real time SLA (few seconds)

Challenge:

- Hadoop MapReduce too slow (> 30 sec)
- Data scanning slow due to constant IO
- Overall job takes > 1 minute

Customer Use Case: Solution

- Utilize existing servers
  Start GridGain data node on every server
- Only put highly utilized files in GGFS
  User controlled caching
- In-Memory MapReduce over GGFS
  Embedded processing
- **Results under 3 seconds**



GridGain

IN-MEMORY COMPUTING

@gridgain

GridGain Systems
www.gridgain.com

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