TurboRush for Hive

• Hadoop delivers data scalability
• Hive delivers ease of use
• TurboRush for Hive delivers a performance boost
  – No learning curve – drop in
  – Preserve HiveQL investment
  – Increase workload capacity
  – Utilizing data already in HDFS
SCALE YOUR BIG DATA APPLICATION ACROSS MANY CORES AND MULTIPLE NODES

Hive Instance (0.7.1)

QL Engine

MetaStore

CLI

JDBC

HWI

Distributed DataRush

DataRush

DataRush

DataRush

DataRush

Hadoop Distributed File System
Features

- Invoking DataRush engine in single-node or distributed mode
- Delimited text and RC input/output formats
- Hive built-in functions
- User defined functions and aggregation functions
- Joins and aggregations
- Full insert support
- Modular design
  - Implements AST, semantic analysis, optimizations and execution plan
  - Pluggable into other frameworks
Exclusions

- Transforms - MapReduce scripts
- Arrays and structures
- Lateral Views
- HBase, Sequence and other formats
Pervasive DataRush™ V5

... a patented, parallel dataflow platform that eliminates performance bottlenecks in your data-intensive applications

- **Scalable**: Performance dynamically scales with increased core counts and increased nodes.

- **High Throughput**: Fast, deep analysis of large data sets with no limit on input data size.

- **Cost Efficient**: Maximum performance from commodity multicore servers, SMP systems and clusters.

- **Easy to Implement**: No complex parallel processing issues; visual and API level interfaces.

- **Extensible**: Extensible platform so you remain in control of development.
TPC-H Based Benchmark

- TPC-H 1000x
- 1 TB data size
- 5 EC2 HPC nodes
  - 24 GB mem
  - 1.6 TB disk
  - 16 cores
TPC-H Based Benchmark

- TPC-H 100x
- 100 GB data size
- 2 node cluster
  - 128 GB mem
  - 96 TB disk
  - 48 cores
Current State

• Early Beta Program
  – Seeking early adopters

• Logical Plan Work
  – Being used as basis for next generation DataRush
  – Composition model allows building distributed applications easily
  – Integrated GUI for drag-n-drop workflow development
Next Steps

- Integrate with Next-Gen MapReduce
- Allows integrating dataflow as a first class citizen within a Hadoop-based cluster