An Introduction to GPU-Accelerated Analytics with Kinetica

Cloud Expo
Karthik Lalithraj, Principal Systems Engineer
Evolution of Data Processing and the Power of the GPU
Developed to Identify Terroristic Threats in Real-Time

- Kinetica incubated as a massively parallel computational engine for US Army INSCOM
- Ingests 50+ sources of streaming data producing 200B new records per hour
- Incorporates geospatial and temporal data
- Real-time, actionable threat intelligence
- First high-performance database to leverage the power of GPUs (deployed 2012)
GPU Acceleration Overcomes Processing Bottlenecks

GPUs are designed around thousands of small, efficient cores that are well suited to performing repeated similar instructions in parallel. This makes them well-suited to the compute-intensive workloads required of large data sets.

High performance computing trend to using GPU’s to solve massive processing challenges.

Parallel processing is ideal for scanning entire dataset & brute force compute.

GPU acceleration brings high performance compute to commodity hardware.

4,000+ cores per device in many cases, versus 16 to 32 cores per typical CPU-based device.
## Evolution of Data Processing

<table>
<thead>
<tr>
<th>Simple Reporting</th>
<th>Standard Analytics</th>
<th>Real-time Analytics</th>
<th>Machine Learning</th>
<th>Deep Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>List defaults from customers in the last 3 years.</td>
<td>What is the default rate for customers over a certain age, by region? by income?</td>
<td>What is the risk-profile of this customer up to and including the transactions he made 10 seconds ago?</td>
<td>Given location, buying history, demographic, past-history, past-purchases, what is the likelihood this customer will default?</td>
<td>Deduce from unspecified signals across a wide range of datasets the likelihood this customer will default?</td>
</tr>
</tbody>
</table>

**MOST BENEFIT FROM GPU**

- Simplified Reporting
- Standard Analytics
- Real-time Analytics
- Machine Learning
- Deep Learning
Kinetica: GPU-Accelerated, In-Memory Analytics Database

1. EXTREME ANALYTIC PERFORMANCE
   - GPU-accelerated database operations

2. HISTORIC AND PREDICTIVE INSIGHTS
   - In-memory, columnar, no tuning, indexing
   - Converge ML, Deep Learning, NLP, and BI
   - Distributed visualization pipeline built in

3. RICH, CONTEXTUAL INSIGHTS
   - Native GIS and IP-address object support
   - Real-time streaming analytics

4. INNOVATION WITHOUT DISRUPTION
   - Deep IT integration: Hadoop, Spark, Tableau, databases, EDW, Kibana
   - Predictable scale out for data ingestion and querying
Kinetica’s Core Capabilities
Kinetica: A Distributed, In-Memory Database

- GPU-accelerated database operations
- Natural language processing based full-text search
- Native GIS and IP-address object support
- Real time data handlers to ingest structured and unstructured data
- No typical tuning, indexing, and tweaking
- Predictable scale out for data ingestion and querying
- Deep integration with open source and commercial frameworks and applications: Hadoop, Spark, NiFi, Accumulo, H2O, Tableau, Kibana and Caravel
- Distributed visualization pipeline built in
Kinetica: Unique Strengths & Capabilities

Fast, Distributed, OLAP Engine for Fast Moving, Large Scale Data

Kinetica is designed to take advantage of the parallel processing nature of the GPU. It delivers low-latency, high performance analytics on large data sets, and makes streaming data available for query in real-time.

Native Geospatial & Visualization Pipeline

Native visualization pipeline makes it easier to work with large geospatial data sets. Ideal for IoT use-cases, and powering geospatial applications.

Interactive Location-Based Analytics

In-Database Analytics

Converged AI and BI

User-defined functions (UDFs) allow for distributed custom compute directly from within the database. In-database analytics on the GPU open the way for AI and BI to run on the same platform.
Architecture Overview
Kinetica: Core

ANALYTICS DATABASE ACCELERATED BY GPUs

Columnar in-memory database

Data available much like a traditional RDBMS... rows, columns

Data held in-memory; persisted to disk

Interact with Kinetica through its native REST API, Java, Python, JavaScript, NodeJS, C++, SQL, etc... as well as with various connectors

Native GIS & IP address object support

VERY FAST: Ideal for OLAP workloads

Typical hardware setup: 256GB - 1TB memory with 2-4 GPUs per node.
Real-Time Data Handlers for Structured & Unstructured Data

- **Reliable, Available and Scalable**
  - Disk based persistence
  - Add nodes on demand
  - Data replication for high availability
  - Scale up and/or out

- **Performance**
  - GPU Accelerated (1000’s Cores per GPU)
  - Ingest Billions of records in minutes
  - Ultra low latency query performance

- **Massive Data Sizes**
  - 100’s of Terabytes Scale
  - Billions of entries

- **Connectors**
  - ODBC/JDBC
  - Restful Endpoints
  - Rich API’s
  - Standard Geospatial Capabilities

- **Run Anywhere**
  - On premise or On Cloud

- **Hardware Partners**
  - Commodity Hardware
Additional Features
AI & BI on One GPU-Accelerated Database

- **Business Intelligence (BI)**
  - Power BI
  - Tableau
  - MicroStrategy

- **Custom Applications**
  - ODBC / JDBC
  - Native REST API
  - WMS

- **High Performance Analytics Database**
  - SQL
  - UDF
  - High Fidelity Geospatial Pipeline

- **Machine Learning & Deep Learning**
  - Caffe
  - BIDMach
  - TensorFlow

- **GPU-Accelerated Data Science**
  - Predictive Models: e.g. Risk Management, Sales Volume, Fraud.

- **Data Scientists / Developers**
  - SQL
  - Data Scientists
  - Developers

- **Business Users**

---

15
Kinetica Reveal

WEB-BASED INTERACTIVE VISUALIZATION FRAMEWORK – INCLUDED WITH KINETICA

• Visualize and analyze billions of data points in human interactive time/instantly
• Make real time dashboards by picking, dragging and dropping an assortment of analytical widgets
• Location based analytics integrate with all major map providers. Choose your preferred base maps: Bing, ESRI, Google, MapBox, open street maps.
• Built in security; multi-level security; permission-based widgets, views, and dashboards
Kinetica High-Speed Geospatial Pipeline

ON DEMAND SCALE OUT +

GEO WEB SERVER
- WMS Layers
- Video
- REST API
  - SQL via ODBC
  - JS
  - Java, C++
  - Python
  - More ...

BI APPLICATIONS
- KINETICA REVEAL
- OTHER DASHBOARDS
  - Tableau
  - Caravel / Kibana
  - ESRI Web / Desktop
  - Etc.

IN-BUILT GEOSPATIAL WEB SERVER REMOVES NEEDS FOR MIDDLE TIER AND ENABLES FAST RENDERING OF WMS OVERLAYS

HIGH VOLUME STREAMING IoT DATA

PARALLEL INGEST

GPU-ACCELERATED GEOSPATIAL DATABASE

MILLISECONDS

IN- BUILT GEOSPATIAL WEB SERVER REMOVES NEEDS FOR MIDDLE TIER AND ENABLES FAST RENDERING OF WMS OVERLAYS
VRAM Boost Mode
TIERED MEMORY MANAGEMENT FOR CAPACITY AND PERFORMANCE

• Features
  • “Pinning” data in very fast GPU Video RAM (VRAM) for lightning fast query performance

• Benefits
  • 100x faster, 1/10 the hardware
  • Prioritize data for ultra low latency
  • Performance with scale via tiered memory management
  • Richer insights with spatial analytics
Full-Text Search

Kinetica includes powerful text search functionality, including:

- Exact Phrases
- Boolean – AND / OR
- Wildcards
- Grouping
- Fuzzy Search (Damerau-Levenshtein optimal string alignment algorithm)
- N-Gram Term Proximity Search
- Term Boosting Relevance Prioritization

“Rain Tire” ~5
"Union Tranquility" ~10
[100 TO 200]
Demo’s