

# Big Data Solutions for Business Analytics

## A Complete Turnkey Solution for Apache Hadoop




# PHAT-DATA<sup>40G</sup>

### Key Features

- A complete turnkey solution bundled with supporting software stack
- Future proof solution equipped with 40G Ethernet out of the box
- High Performance Hadoop Platform featuring an elegant user interface
- Integrated ultra-low latency PCIe Solid-State Drives
- Ease of management with minimal labor costs since the expertise is built into the system
- Monitoring, fail-over and recovery, alerts, security and authentication
- Supports multiple file systems and data sources including HDFS, GPFS and any file system with HDFS plug-in
- Self-monitoring and troubleshooting capabilities
- Support for up to 300 job trackers across a shared pool of resources
- Application and job level predictability via sophisticated job scheduler to guarantee SLA requirements
- Support for homogenous and heterogeneous network environments

### HADOOP USE CASES

Use Case	Application	Industry	Application	Use Case
ADVANCED ANALYTICS	Social Network Analysis	Web	Clickstream Sessionization	DATA PROCESSING
	Content Optimization	Media	Clickstream Sessionization	
	Network Analysis	Telco	Mediation	
	Loyalty & Promotions Analysis	Retail	Data Factory	
	Fraud Analysis	Financial	Trade Reconciliation	
	Entity Analysis	Federal	SIGINT	
	Sequencing Analysis	Bioinformatics	Genome Mapping	

### The Big Data problem - is it really that Big?

Organizations are facing a very critical challenge today due to the sheer volume of data being generated every year. This data, whether structured or unstructured, can offer invaluable insights and information about the internal details of an organization. Current data analysis methods prove largely inadequate to gain any valuable insight from such large, multi-terabyte, and even petabyte, datasets. Corporate sectors such as media, healthcare, manufacturing, retail, etc all are facing the challenge of analyzing this amount of data. Enter PHAT-Data.

AMAX's PHAT-Data (Peta-Scale Hadoop Analytics Technology) provides a complete turnkey solution for Apache Hadoop. The PHAT-Data offers a highly scalable, tunable and easy to deploy platform bundled with Apache Hadoop to solve Big Data acquisition, storage and analysis problems, and is thoroughly tested and certified production-ready to be deployed on large scale. The PHAT-Data turnkey solution includes a complete software stack of Apache Hadoop, configured according to your application specification, with a dedicated Hadoop engineering team standing by to help tune your cluster for optimum efficiency and maximum performance. Turn on the power button and you are good to go.

### Simplicity to the Core

Hadoop is an ensemble where each piece including Sqoop, Flume, Pig and Hive etc. are integral pieces of puzzle. Being tuned to your needs, AMAX PHAT-Data is simple to deploy with a click of a button. The entire provisioning process is automated and customized based on your requirements.

### Performance Tuning Capability

Big Data solutions using Hadoop consist of an ensemble of tools which require precise tuning and configuration for optimum efficiency and maximum performance. For example, a Hadoop based database utilizes HBase - a real-time database which can use compression to minimize network overhead – which requires specific tuning to extract the full potential. AMAX understands your application and tunes each turnkey cluster so that guesswork is taken out of the equation.

### Open Architecture

As a part of our commitment to an open architecture design philosophy, AMAX PHAT-Data does not include any proprietary components. The entire solution offers 100% compatibility with Java based MapReduce applications and support heterogeneous environments and open standards right out of the box.

### High Availability

The Hadoop NameNode can become single point of failure if the cluster is not configured for high availability. The AMAX PHAT-Data includes 2 name nodes configured with DRBD/Heartbeat or RSF-1 in master-slave configuration with a dedicated heartbeat network to address this issue. The auxiliary NameNode automatically takes over in case of a failure. Our solution promises no downtime and absence of a single point of failure. The MapReduce engine offers automatic job recovery without the need to restart in case of JobTracker/TaskTracker failure.

### Ready for Production Environment

Each and every component selected in the PHAT-Data cluster solution passes through AMAX's stringent 5 stage quality test which includes individual component tests as well as a complete system integration test. Our ISO 9001 certified manufacturing facility and 3 decades of engineering expertise promises utmost care at every level irrespective of the volume.

## System Configuration

### Standard configuration includes:

- 5x Nodes(NameNode/DataNode)
- 4x PCI-E SSD per node
- 1x 40GbE Switch
- 1x 1GbE Switch

### NameNode/DataNode:

- 2x Intel®Xeon® E5-2600 processor series
- 128GB DDR3 1333Mhz ECC Reg.RAM(Up to 512GB)
- 4x PCI-E 980GB SSD Drive
- 1x 2.5" 120GB SSD drive(Up to 16 drive bays)
- 1x Dual-port 40 Gigabit Ethernet Adapter
- 4x Intel® i350 Gigabit Ethernet
- 1x Intel® RMM4 with complete IPMI 2.0 support

### Network Fabric:

- 1x High performance and low latency 40GbE switch with Virtual Protocol Interconnect® (VPI) support
- 1x 1GbE switch dedicated for management

### Flash Based Storage:

- PCIe Solid-State-Drive
- Low latency – provide faster responses to host requests, and accelerates enterprise applications
- Integrated ASIC-based architecture - Integrates flash subsystem and PCIe technology in a power-efficient, single-ASIC solution to provide high application performance, while reducing server CPU and memory footprints for the lowest cost solution
- Low-profile PCIe form factor
- Low-power design

### Software Components:

- **Orchestrator** – Management, Administration, Automation, Provisioning, Security (MAAPS)
  - Management - elegant, intuitive, yet powerful administration of your Hadoop cluster
  - Automation - automated monitoring, reporting, failover, and performance tuning
  - Analytics - our web UI puts all the power of Hadoop analytics in your browser
  - Security - ensure your cluster and analytics have flexible, strongly enforced safety policies
- **Unstructured Data Accelerator(UDA)** – accelerates Hadoop network and improves the scaling of Hadoop clusters executing data analytics intensive applications
  - More than doubles the data processing throughput and reduces total job execution time by half per Hadoop node
  - It is designed to scale with larger datasets to provide similar or better performance benefits

- Leverages world's fastest interconnect that supports 40Gb/s InfiniBand or Ethernet fabric
- Lossless scalable fabric solution
- **Apache Hadoop** – Open source software framework for scalable, reliable distributed computing.
  - Hadoop Common
  - Hadoop Distributed File System (HDFS)
  - Hadoop MapReduce
- **Apache HBase** – Database for real-time, random access to data stored using Hadoop.
  - Apache Hive – Data warehouse using Hadoop for data summarization, ad-hoc queries and analysis.
  - Apache Pig – High level language for expressing data analysis.
  - Apache Sqoop – SQL to RDBMS integration tool.
  - Apache Flume – Distributed, reliable service to aggregate log data.
  - Apache Whirr – A set of libraries to run cloud services.

### Supported Operating Systems

- Red Hat Enterprise Linux 5 and 6
- SuSe Enterprise Linux Server 10 and 11
- CentOS 5 and 6
- Ubuntu Server 10,11
- Microsoft Windows Server 2008 R2

