EMC GREENPLUM

# ADVANCED CYBER ANALYTICS WITH GREENPLUM DATABASE

## Massively Scalable Data Warehousing Meets Large-Scale Analytics Processing

### GREENPLUM IN GOVERNMENT

Greenplum is working with a number of leading government organizations to consolidate their data and perform the complex, mission-critical analysis necessary to support their overall mandates and charters.

Greenplum's database is being deployed to solve some of the most challenging analytical and data processing applications including:

- Fraud detection and cyber security
- Compliance and regulatory analysis
- Energy consumption and carbon footprint managements
- Large-scale trend analysis

### THE GREENPLUM STORY

Data warehousing and analytics innovator Greenplum was acquired by EMC in July 2010, becoming the foundation of the new EMC Data Computing Division. Greenplum's software and appliance products deliver the big data analytics platform of the future, enabling actionable insight from vast and diverse datasets.

## COMPUTER NETWORK DEFENSE STRATEGY REQUIRES CYBER ANALYTICS

Today's "cyber domain" is growing rapidly in both size and complexity with the proliferation of networked communication devices, cloud-based compute services, and network access points. This creates several new challenges in the effort to build a sophisticated Computer Network Defense (CND) strategy to protect, monitor, analyze, detect, and respond to unauthorized activity within an enterprise's network(s). As part of such a strategy, organizations and government agencies are performing cyber analytics as the means to implement both predictive defense and adaptive approaches towards cyber security.

Sophisticated attackers may go undetected for a long period of time and are the ones responsible for the majority of the damage. These types of exploits are referred to as "slow and low" attacks. Unfortunately, current cyber security devices generally do a poor job of detecting these attacks. Exploitations such as zero-day exploits, intrusion prevention/detention system (IPS/IDS) evasion, and stealth network reconnaissance usually span days to months and sometimes even years making it difficult to recognize any type of pattern or signature.

### THE CHALLENGES

To halt these types of attacks and exploits, vast amounts of data need to be first collected and then analyzed. This data needs to become actionable in order to understand past and current events, and how to prepare for future events. However, to transform and correlate data into actionable cyber analytics results requires overcoming three challenges:

- **Big Data** – How can terabytes and petabytes of data be transformed into information that enables vital discoveries and timely decisions?
- **Sophisticated Requirements** – What is available to support adaptive rule engines and models that will evolve faster than those of adversaries?
- **Fast Performance** – How can analysts be enabled to get results quickly (in a matter of seconds) from their complex queries?

As the cyber domain continues to grow, more and more cyber data is being collected inreal-time by log capture, commercial IDS/IPS products, homegrown network sensors, and other sources. This cyber data quickly becomes overwhelming as it is captured at numerous points and in various formats throughout the network. As this happens, it becomes apparent that the major inhibitors to successfully accomplishing enterprise-wide cyber analytics is the requirement to handle large amounts of data, while also having the ability to apply sophisticated (and adaptive) rules engines, models, and scoring techniques in a near real-time window. In other words, there is a requirement for a massively scalable data warehouse that will act as the foundation for advanced analytics, having the ability to ingest cyber data at incredible rates to produce actionable steps.

## CURRENT DATABASE ARCHITECTURES ARE LIMITED

### ONLINE TRANSACTION PROCESSING (OLTP) VS. MASSIVELY PARALLEL PROCESSING (MPP)

Greenplum database utilizes a shared-nothing, massively parallel processing (MPP) architecture that has been designed for analytical processing. Most of today's general-purpose relational database management systems are designed for Online Transaction Processing (OLTP) applications. OLTP transaction workloads require quick access and updates to a small set of records. This work is typically performed in a localized area on disk, with one or a small number of parallel units. Shared-everything architectures, in which processors share a single large disk and memory, are well suited to OLTP workloads (Figure 1). Shared-disk architectures, such as Oracle and the underlying databases for today's security information and event management (SIEM) products are quickly overwhelmed by the full table scans, multiple complex table joins, sorting, and aggregation operations against vast volumes of data. OLTP architectures are not designed for the levels of parallel processing required to execute complex analytical queries, and tend to bottleneck as a result of failures of the query planner to leverage parallelism, lack of aggregate I/O bandwidth, and inefficient movement of the data between nodes.

To transcend these limitations, Greenplum built a shared-nothing MPP database, designed from the ground up for data warehousing and large-scale analytics processing. The Greenplum database shared-nothing architecture separates the physical storage of data into small units on individual segment servers each with a dedicated, independent, high-bandwidth channel connection to local disks (Figure 2). The segment servers are able to process every query in a fully parallel manner, use all disk connections simultaneously, and efficiently flow data between segments as query plans dictate. Because Greenplum's shared-nothing database automatically distributes data and makes query workloads parallel across all available hardware, it dramatically outperforms general-purpose database systems on analytical workloads.

SHARED-DISK
(IE: ORACLE RAC)

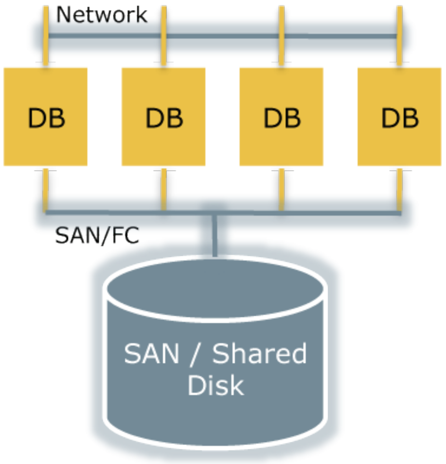

Figure 1: Traditional OLTP Database Shared-Everything Architecture

## NEW CAPABILITIES IN DATA INTENSIVE COMPUTING

The Pacific Northwest National Laboratory (PNNL), a U.S. Department of Energy laboratory, and Greenplum are working together to deliver innovative cyber analytics and data-intensive computing solutions to the public sector.

The work at PNNL incorporates multiple streams of data to benefit both human and automated analysis. Cyber data such as host, network, and information from the World Wide Web are being fused with human-oriented data such as psycho-social, political, and cultural information to form a more complete understanding of adversaries, motives, and social networks.

This and numerous other cyber analytics projects are ongoingat PNNL, including deep forensics, critical infrastructure protections, SCADA security, and insider threat protection.

USASPENDING.GOV
Government spending at your fingertips

## TRANSPARENCY DELIVERED

The USAspending.gov site is an Internet-scale data and analytics platform that leverages the industry's most cutting-edge database, distributed systems, virtualization and cloud computing technologies. The core database engine, powered by Greenplum, receives fine-grained spendingdetails and aggregates and distills this information into a form that can easily be visualized and explored by the public. It fully utilizes the parallelism of dozens of compute cores to perform SQL-based aggregation and analysis, and employs both row-oriented and column-oriented processing as well as powerful in-database compression to maximize performance. The results of analysis and aggregation flow to a scalable frontline tier of Greenplum Database instances that are tuned to perform complex slicing and dicing of this data to large volumes of users concurrently.

All of these components arerunning within Nebula, an innovative, efficient and highly scalable private federal cloudcomputing platform developedby NASA to make its vastquantities of data available tothe public.

SHARED-NOTHING
(IE: GREENPLUM)

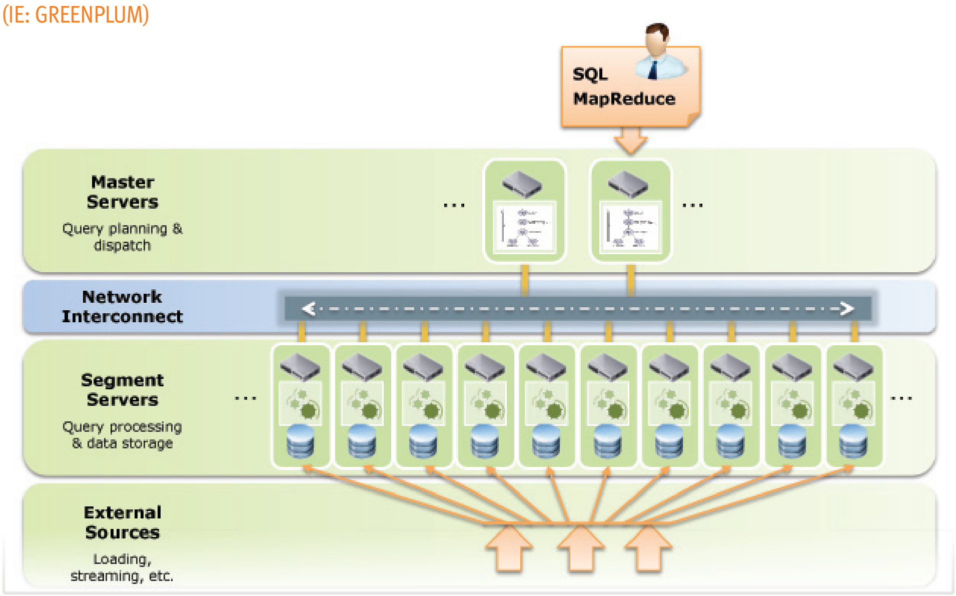

Figure 2: Greenplum's MPP Shared-Nothing Architecture

## THE GREENPLUM SOLUTION

Organizations and government agencies employ Greenplum's database software as the data-warehousing platform for large-scale Cyber Analytics. Greenplum's database software allows a cluster of commodity servers and storage to operate as a single data warehouse supercomputer, automatically partitioning data and parallelizing queries to achieve performance tens or hundreds of times of faster than traditional cyber tools or relational databases. With production data warehouses ranging from several terabytes to well over the petabyte threshold, Greenplum easily addresses the scalability requirement for cyber analytics. Additionally, Greenplum's Massively Parallel Processing (MPP) shared-nothing architecture allows for data ingest rates of multiple TBs/hour, while not hindering the analytics anduser queries.

Greenplum's database software is more than a typical data warehouse. With Greenplum Database 4.0, cyber analysts and data analysts are able to use common analytical tools like SQL, C/C++, Java, Python, Perl, R, SAS, and many others to interrogate the data. Analysts can also use Greenplum's built-in functionality and support of MapReduce.

Historically, cyber forensics and cyber analytics have been limited to after-the-fact detection rather than preventive defense. However, if an analyst or automated systems are able to recognize adverse activity as it is occurring, there is the possibility of preventing it. An example from industry is credit card fraud detection, where Greenplum databases are often used as the underlying platform to recognize, in real time, that a transaction does not meet a user's profile, and then send an alert or prevent the transaction from being completed.

## SUMMARY

When Greenplum's database is used for cyber analytics, it allows organizations to perform pattern analysis, anomaly detection, and other operations to accomplish deep cyber forensic analysis and uncover malicious network activity. Greenplum can help discover non-obvious relationships on the network by quickly correlating and analyzing cyber data. Some of the key benefits of implementing Greenplum database for cyber analytics are:

- Scalability – from tens of terabytes to multiple petabytes
- Fast queries results and load times run against a massively parallel processing database
- Holistic view of enterprise-wide cyber security posture
- Cost-effective acquisition and operational costs
- Advanced “in-database” analytics
- Support of row- and column-oriented database tables
- Compressions
- Better analytic performance
- Support of MapReduce
- Support for PMML (predictive modeling mark-up language)
- Data Mining Models
- Sophisticated text analytic capabilities
- Flexibility to implement Greenplum in a private cloud or on dedicated servers
- Self-service provisioning of databases/datamarts for dedicated analysis
- Web interface for data analyst collaboration

A successful CND is more than network sensors positioned throughout a network and then locally queried and analyzed by a SIEM product. CND requires an orchestrated approach with the ability to have a real-time holistic view of events and threats. Cyber analytics is about pattern recognition and real-time analytics. This requires a database built specifically for data warehousing and analytical processing, such as Greenplum database.

## GREENPLUM PRODUCTS

### Greenplum Database

Greenplum Database, a major release of Greenplum’s industry-leading massively parallel processing (MPP) database product, is the culmination of more than seven years of development. Already regarded as the most scalable mission-critical analytical database and in use at more than 200 leading enterprises worldwide, in this new release, Greenplum Database raises the bar with numerous powerful features and enhancements.

### Greenplum Chorus Software

Greenplum Chorus Software is a new class of software that empowers people within an enterprise to more easily collaborate and derive insight from their data. As the first commercial Enterprise Data Cloud platform, it provides the key services necessary to realize the benefits of private cloud computing techniques and social collaboration for enterprise data warehousing and analytics.

### Greenplum HD Product Family

The EMC Greenplum HD product family enables you to take advantage of big data analytics without the overhead and complexity of a project built from scratch.

### Greenplum Data Computing Appliance Family

This family of advanced “all-in-one” appliances delivers a fast loading, highly scalable, parallel computing platform for next-gen data warehousing and analytics.

## CONTACT US

To learn more about how EMC Greenplum products, services and solutions can help you realize big data opportunities visit us at www.greenplum.com.

**ABOUT EMC GREENPLUM AND THE DATA COMPUTING PRODUCTS DIVISION** EMC’s new Data Computing Products Division is driving the future of data warehousing and analytics with breakthrough products including EMC Greenplum Data Computing Appliance Family, EMC Greenplum Database, EMC Greenplum Database Community Edition, and EMC Greenplum Chorus— the industry’s first Enterprise Data Cloud platform. The division’s products embody the power of open systems, cloud computing, virtualization, and social collaboration—enabling global organizations to gain greater insight and value from their data than ever before possible.

**EMC Greenplum**
Division Headquarters
1900 South Norfolk Street San Mateo, CA 94403
Tel: 650-286-8012
www.greenplum.com