

**DATA MINING AND BIG DATA ANALYTICS
(18MCA52C)
UNIT V**

FACULTY

Dr. K. ARTHI MCA, M.Phil., Ph.D.,
Assistant Professor,
Postgraduate Department of Computer Applications,
Government Arts College (Autonomous),
Coimbatore-641018.

What is Data?

- ▶ The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.



What is Big Data?

- ▶ Big Data is also data but with a huge size.
- ▶ Big Data is a term used to describe a collection of data that is huge in volume and yet growing exponentially with time.
- ▶ In short such data is so large and complex that none of the traditional data management tools are able to store it or process it efficiently.



Examples Of Big Data

Social Media

- ▶ The statistic shows that 500+terabytes of new data get ingested into the databases of social media site Facebook, every day.
- ▶ This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.

The image shows the Facebook logo, which consists of the word "facebook" in a blue, lowercase, sans-serif font. The logo is centered within a white rectangular box that has a subtle gradient and a slight drop shadow, making it stand out against the background. The background of the slide features abstract green and white geometric shapes on the right side.

Types Of Big Data

Big Data' could be found in three forms:

- ▶ Structured
- ▶ Unstructured
- ▶ Semi-structured



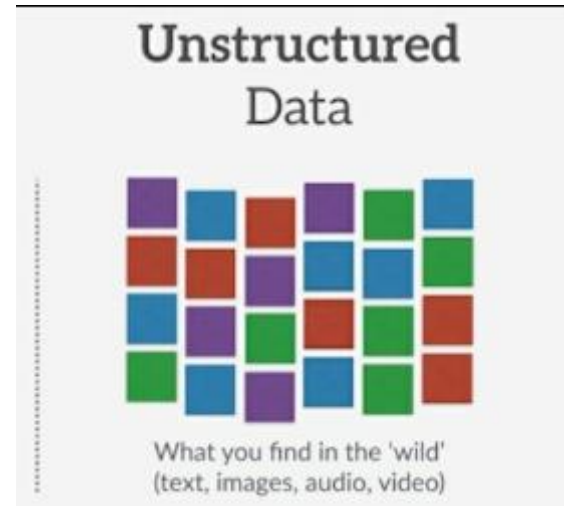
Structured

- ▶ Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.



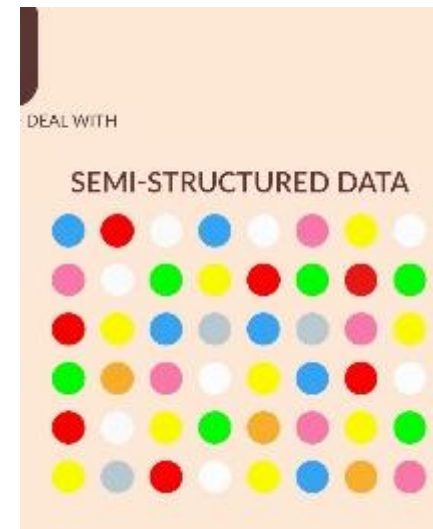
Unstructured

- ▶ Any data with unknown form or the structure is classified as unstructured data.
- ▶ In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it.
- ▶ A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.



Semi-structured

- ▶ Semi-structured data can contain both the forms of data.
- ▶ We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS.



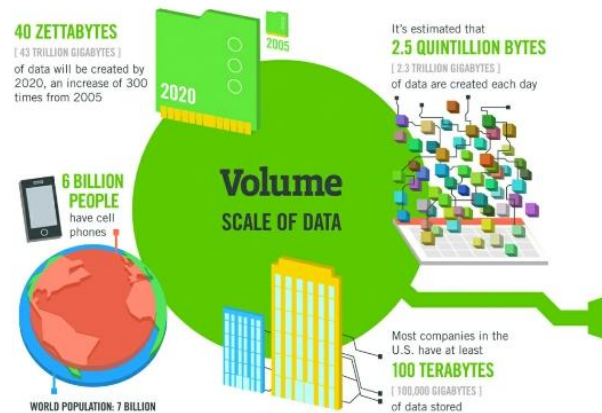
Characteristics Of Big Data

- ▶ Volume
- ▶ Variety
- ▶ Velocity
- ▶ Variability



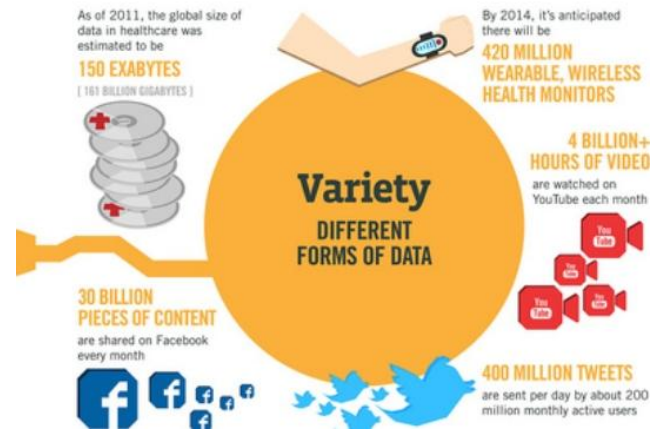
Volume

- ▶ The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data.



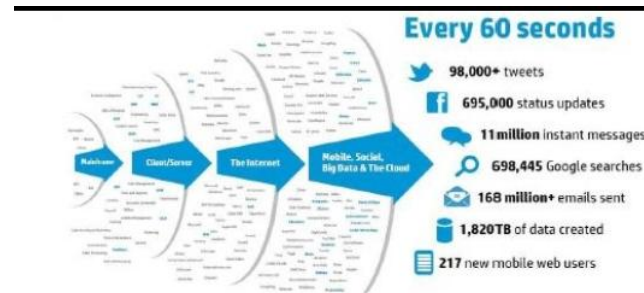
Variety

- ▶ Variety refers to heterogeneous sources and the nature of data, both structured and unstructured.
- ▶ Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications.



Velocity

- ▶ The term 'velocity' refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.



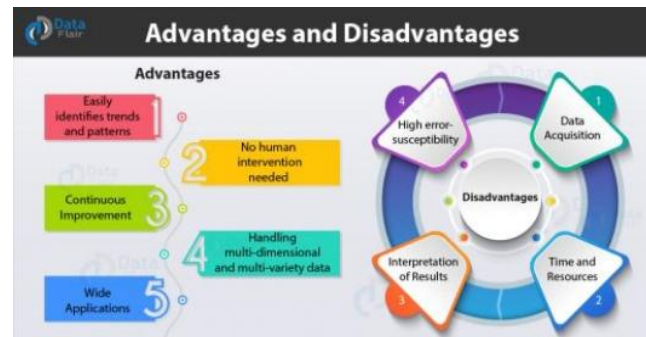
Variability

- ▶ This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.



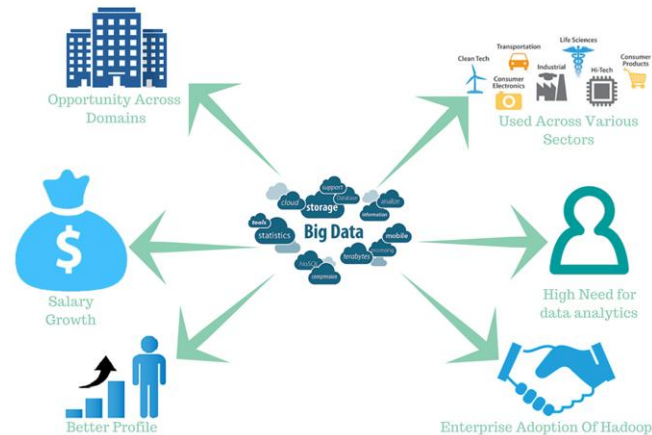
Benefits of Big Data Processing

- ▶ Businesses can utilize outside intelligence while taking decisions
- ▶ Improved customer service
- ▶ Early identification of risk to the product/services, if any Better operational efficiency



Importance of Big Data

- ▶ The Big Data analytics is indeed a revolution in the field of Information Technology.
- ▶ Big data has the properties of high variety, volume, and velocity.
- ▶ You can store Tbs of data, pre process it , analyze the data and visualize the data with the help of couple of big data tools.



Main reasons

- ▶ Following are the three main reasons that why Big data is so important and efficient.
- ▶ Cost reduction
- ▶ Faster, better decision making
- ▶ New products and services



Cost reduction

- ▶ Big data technologies such as Hadoop and cloud-based analytics bring significant cost advantages when it comes to storing large amounts of data



Faster, better decision making

- ▶ With the speed of Hadoop and in-memory analytics, combined with the ability to analyze new sources of data, businesses are able to analyze information immediately and make decisions based on what they've learned.



New products and services

- ▶ With the ability to gauge customer needs and satisfaction through analytics comes the power to give customers what they want.



Real-time Benefits of Big Data Analytics:

The use of Big Data analytics is very flexible to another fields as well. With the use of big data a lot there has been an enormous growth in multiple industries. Some of them are

- ▶ Banking
- ▶ Technology
- ▶ Consumer
- ▶ Manufacturing



IT for IT

- ▶ Hadoop-based log analytics has become a common use, but they doesn't mean it's deployment is as widespread as it should be.
- ▶ Log analytics is actually a pattern than IBM established after Working with a number of companies, initially in FSS.
- ▶ We've since seen this use case come up across all industries, for that reason, we'll call this pattern IT for IT.
- ▶ Big data-enriched IT for IT helped clients to Gain better insight into how their systems are running, and when and how things break down.

DATA CENTER

WHAT IS A DATA CENTER

Data centers provide a range on information technology services

- ▶ These services include:
 - ▶ Email
 - ▶ Data storage and management
 - ▶ Web hosting
 - ▶ Application Hosting

DATA CENTER BENEFITS

▶ Economies of Scale

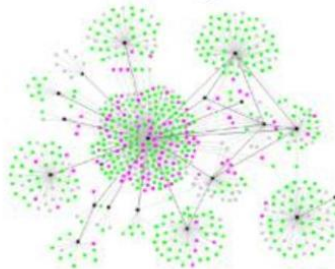
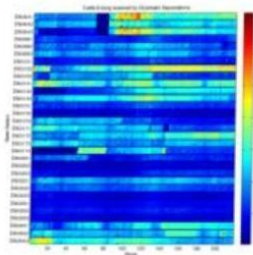
- ▶ Purchasing savings based on large purchases
- ▶ Dedicated IT staff
- ▶ Shared resources

▶ Security

- ▶ Physical
 - ▶ Limited access to servers
 - ▶ Surveillance
- ▶ Virtual
 - ▶ Firewalls
 - ▶ Anti-virus
 - ▶ Password Protection

What is Machine Learning?

- Collection of computational methods to ...
 - Detect hidden patterns in data
 - Create useful predictions about unseen data
 - Decision making under uncertainty
 - Transform raw data into useful knowledge

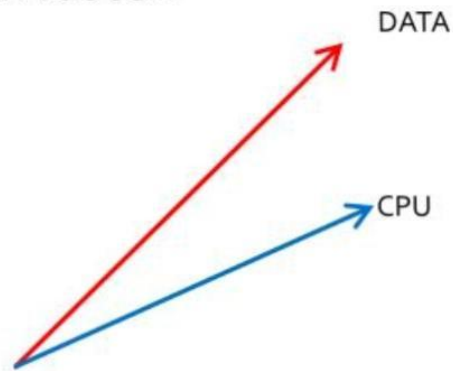


Data Mining, Machine Learning, Statistics

- Facets of the same problem
- Differences in emphasis/terminology
- Historical Evolution of the fields
 - Data Mining: *Database systems, Data Structures*
 - Statistics: *Probability Theory, Mathematics*
 - Machine Learning: *Artificial Intelligence, Pattern Recognition*

Moore's Law to Rescue?

- “data explosion is bigger than Moore's law”
- Computers get faster and cheaper every year but the amount of data that needs to be processed grows even faster.



Use Cases: Government

- Urban Traffic Management
- Energy Grid Management/Optimization,
- Power Generation Management
- Environment Monitoring



- Log Analytics enables IT Operations Analytics for Machine Data
- Correlation of Events is the Key for Added Business Value
- Log Management is complementary to other Big Data Components

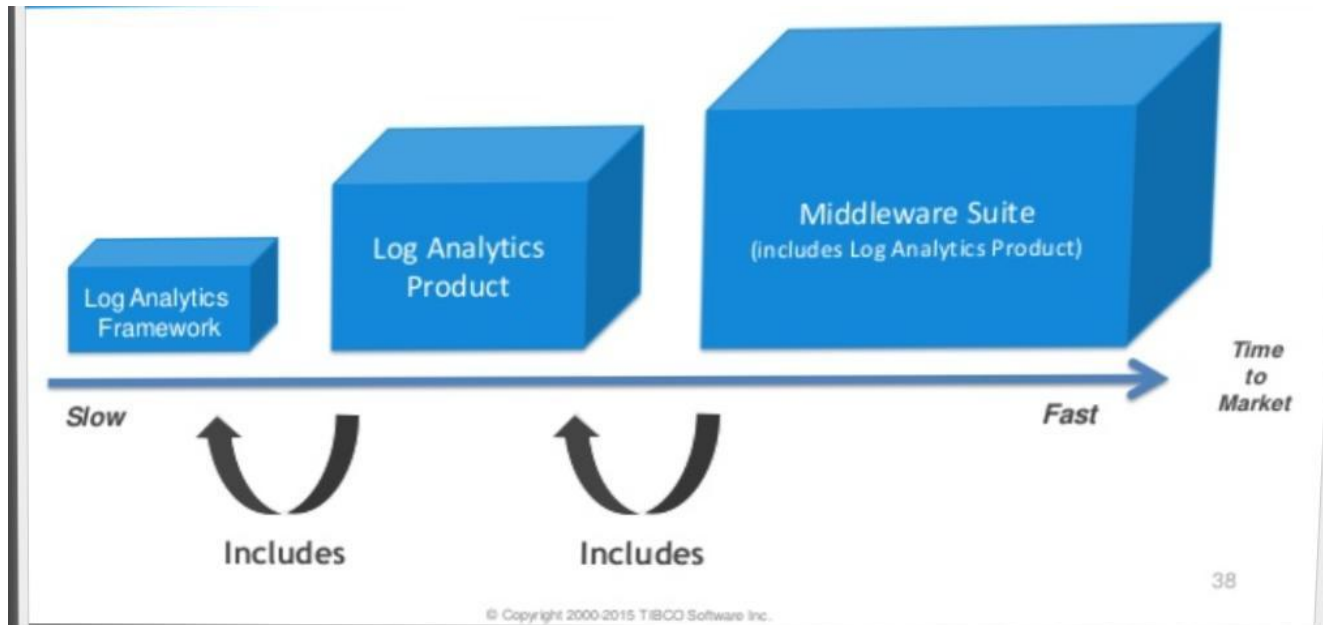
Log Management

- **SaaS** → Easy to setup and use, but cloud cons (not flexible, public cloud)
- **Open Source** → Free and extendable, but coding / config instead of tooling
- **Enterprise** → Most feature-rich and powerful tooling, but more expensive

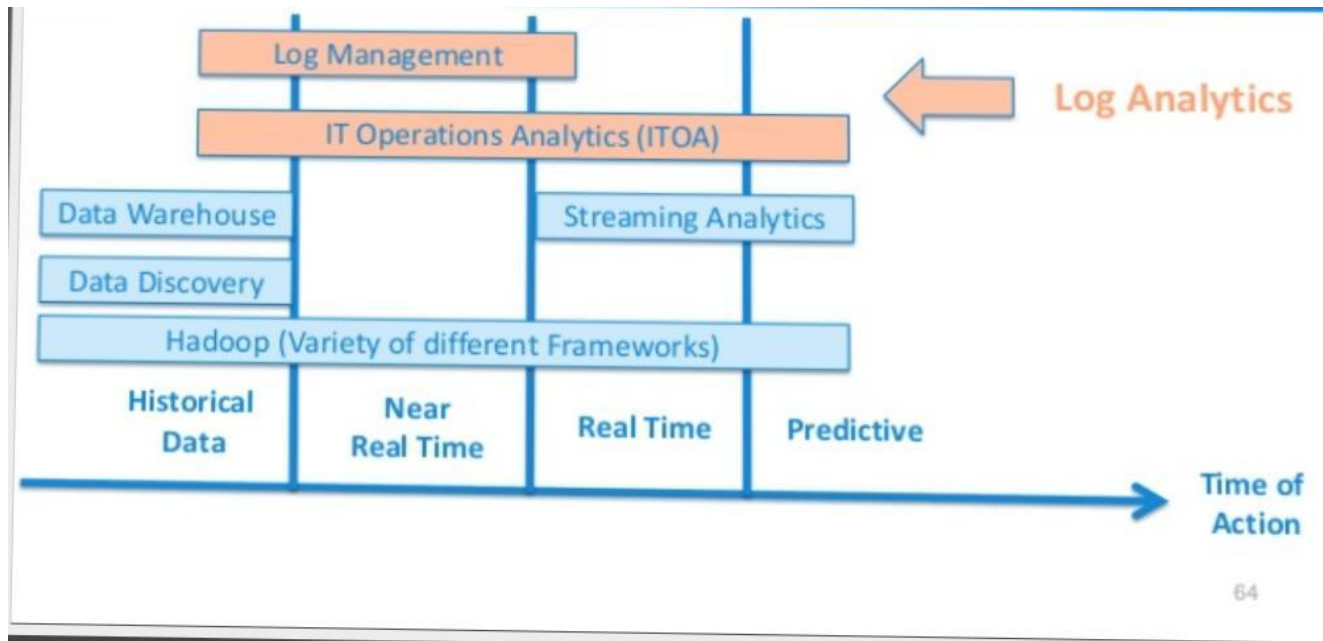
IT Operations Analytics (ITOA)

- Enterprise vendors entering this market these days
 - Extending existing solutions
- Focus on complex correlations, real time processing, predictive monitoring

Alternatives for log analytics:



When to use log analytics



In Simple Language...



“Social media
is people having
conversations online.”

**“Creation of web content, by the
people, for the people”**

SOCIAL MEDIA PLATFORMS



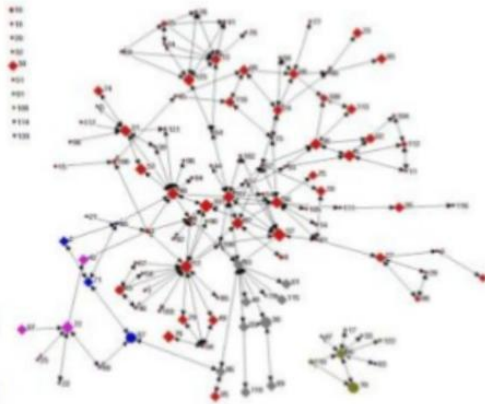
WEB 2.0 TECHNOLOGIES: SOCIAL MEDIA



Social media:
is an umbrella
term that
defines the
various activities
that integrate
technology,
social
interaction, and
the construction
of words,
pictures, videos
and audio.

SOCIAL NETWORK ANALYSIS

- We live in networks all the time: communities, organizations, teams
- There is science to support the understanding of network structure
- The structure of a network provides insights into how the network “works”
- Once you understand the structure, you can make decisions about how to manage the network’s context
- Network analysis tools help you understand the structure



Understanding customer sentiment:

Introduction

- **Two main types of textual information: Facts and Opinions**
- **Most current text information processing methods work with factual information (e.g., web search, text mining)**
- **Sentiment analysis or opinion mining, computational study of opinions (sentiments, emotions) expressed in text**
- **Why opinion mining now? Mainly because of the Web huge volumes of opinionated text.**

SENTIMENT ANALYSIS



happy



sad



angry



disappointed



surprised



proud



in love

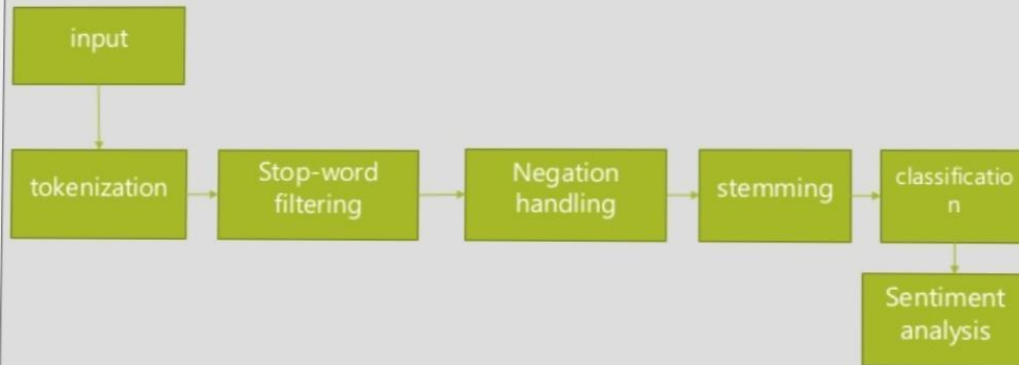


scared

WHAT IS SENTIMENT ANALYSIS?

- The movie was awesome 👍
- The movie was awful 🗑️
- The movie was long 😞

SENTIMENT ANALYSIS WORK FLOW





**Customers do
provide feedback,
lots of feedback**



WHAT?

Is the Challenge

**But How to Process so
Many Customer
Feedbacks?**

HELP




Opinion Mining

- Along with entity, topic and event recognition, opinion mining forms the cornerstone for social web analysis



Healthcare Big Data Use Cases



- 
- ▶ Healthcare organizations are using big data for everything from improving profitability to helping save lives.
 - ▶ Healthcare companies, hospitals, and researchers collect massive amounts of data. But all of this data isn't useful in isolation.
 - ▶ It becomes important when the data is analyzed to highlight trends and threats in patterns and create predictive models.

Genomic Research

- ▶ Big data can play in a significant role in genomic research.
- ▶ Using big data, researchers can identify disease genes and biomarkers to help patients pinpoint health issues they may face in the future.
- ▶ The results can even allow healthcare organizations to design personalized treatments.

Challenges

- ▶ Companies must integrate data coming from different formats and identify the signals that will lead to optimizing maintenance.

Patient Experience and Outcomes

- ▶ Healthcare organizations seek to provide better treatment and improved quality of care—without increasing costs.
- ▶ Big data helps them improve the patient experience in the most cost-efficient manner.
- ▶ With big data, healthcare organizations can create a 360-degree view of patient care as the patient moves through various treatments and departments.

Challenges

- ▶ Improving the patient experience requires a large volume of patient data, some of which could be multi structured data, such as doctor notes or images.
- ▶ Additionally, to analyze patient journeys, path and graph analyses are often needed.

Claims Fraud

- ▶ For every healthcare claim, there can be hundreds of associated reports in a variety of different formats.
- ▶ This makes it extremely difficult to verify the accuracy of insurance incentive programs and find the patterns that indicate fraudulent activity.
- ▶ Big data helps healthcare organizations detect potential fraud by flagging certain behaviors for further examination.

Challenges

- ▶ Claims fraud analytics is a complex process that involves integrating different data sets, analyzing the claims data, and identifying complex fraud patterns.

Healthcare Billing Analytics


- ▶ Big data can improve the bottom line.
- ▶ By analyzing billing and claims data, organizations can discover lost revenue opportunities and places where payment cash flows can be improved.
- ▶ This use case requires integrating billing data from various payers, analyzing a large volume of that data, and then identifying activity patterns in the billing data.

Challenges

- ▶ Sifting through large volumes of data can be complicated, especially when it comes to integrating different data sources.

Telecommunications Big Data Use Cases



- 
- ▶ The popularity of smart phones and other mobile devices has given telecommunications companies tremendous growth opportunities.
 - ▶ But there are challenges as well, as organizations work to keep pace with customer demands for new digital services while managing an ever-expanding volume of data.

Optimize Network Capacity

- ▶ Optimal network performance is essential for a telecom's success. Network usage analytics can help companies identify areas with excess capacity and reroute bandwidth as needed.
- ▶ Big data analytics can help them plan for infrastructure investments and design new services that meet customer demands.
- ▶ With new insights, telecoms are able to maintain customer loyalty and avoid losing revenue to competitors.

Challenges

- ▶ In addition to creating complex models of relationships between network services and customers, network usage analytics requires analyzing a high volume of call detail records.

Telecom Customer Churn

- ▶ By analyzing the data telecoms already have about service quality, convenience, and other factors, telecoms can predict overall customer satisfaction.
- ▶ And they can set up alerts when customers are at risk of churning—and take action with retention campaigns and proactive offers.

Challenges

- ▶ This use case requires analyzing past and current data to create a new model to predict churn, which can be done with time-series and relational analytics to identify patterns and behavior.
- ▶ Graph analytics helps identify relationships between customers who have recently churned and current customers who may be more likely to churn because they know someone who has churned.

New Product Offerings


- ▶ Unstructured sensor and historical data can be used to optimize oil well production.
- ▶ By creating predictive models, companies can measure well production to understand usage rates.
- ▶ With deeper data analysis, engineers can determine why actual well outputs aren't tallying with their predictions.

Challenges

- ▶ This use case involves analyzing a large volume of data.
- ▶ Complex algorithms are also needed to identify the curve shape associated with that data to identify trends.

Financial Services Big Data Use Cases



- 
- ▶ Forward-thinking banks and financial services firms are capitalizing on big data.
 - ▶ From capturing new market opportunities to reducing fraud, financial services organizations have been able to convert big data into a competitive advantage.

Fraud and Compliance

- ▶ When it comes to security, it's not just a few rogue hackers.
- ▶ The financial services industry is up against entire expert teams.
- ▶ While security landscapes and compliance requirements are constantly evolving. Using big data, companies can identify patterns that indicate fraud and aggregate large volumes of information to streamline regulatory reporting.

Challenges

- ▶ Collecting and aggregating disparate data sources can be difficult.

Anti-Money Laundering

- ▶ Financial services firms are under more pressure than ever before from governments passing anti-money laundering laws.
- ▶ These laws require that banks show proof of proper diligence and submit suspicious activity reports.
- ▶ In this extraordinarily complicated arena, big data analytics can help companies identify potential fraud patterns.

Challenges

- ▶ This use case requires analyzing large volumes of transaction data (which can include structured and multi-structured data) and then identifying complex AML transactions.
- ▶ In addition, graph analytics will reveal the hidden relationships.

Financial Regulatory and Compliance Analytics

- ▶ Financial services companies must be in compliance with a wide variety of requirements concerning risk, conduct, and transparency.
- ▶ At the same time, banks must comply with the Dodd-Frank Act, Basel III, and other regulations that require detailed reporting.

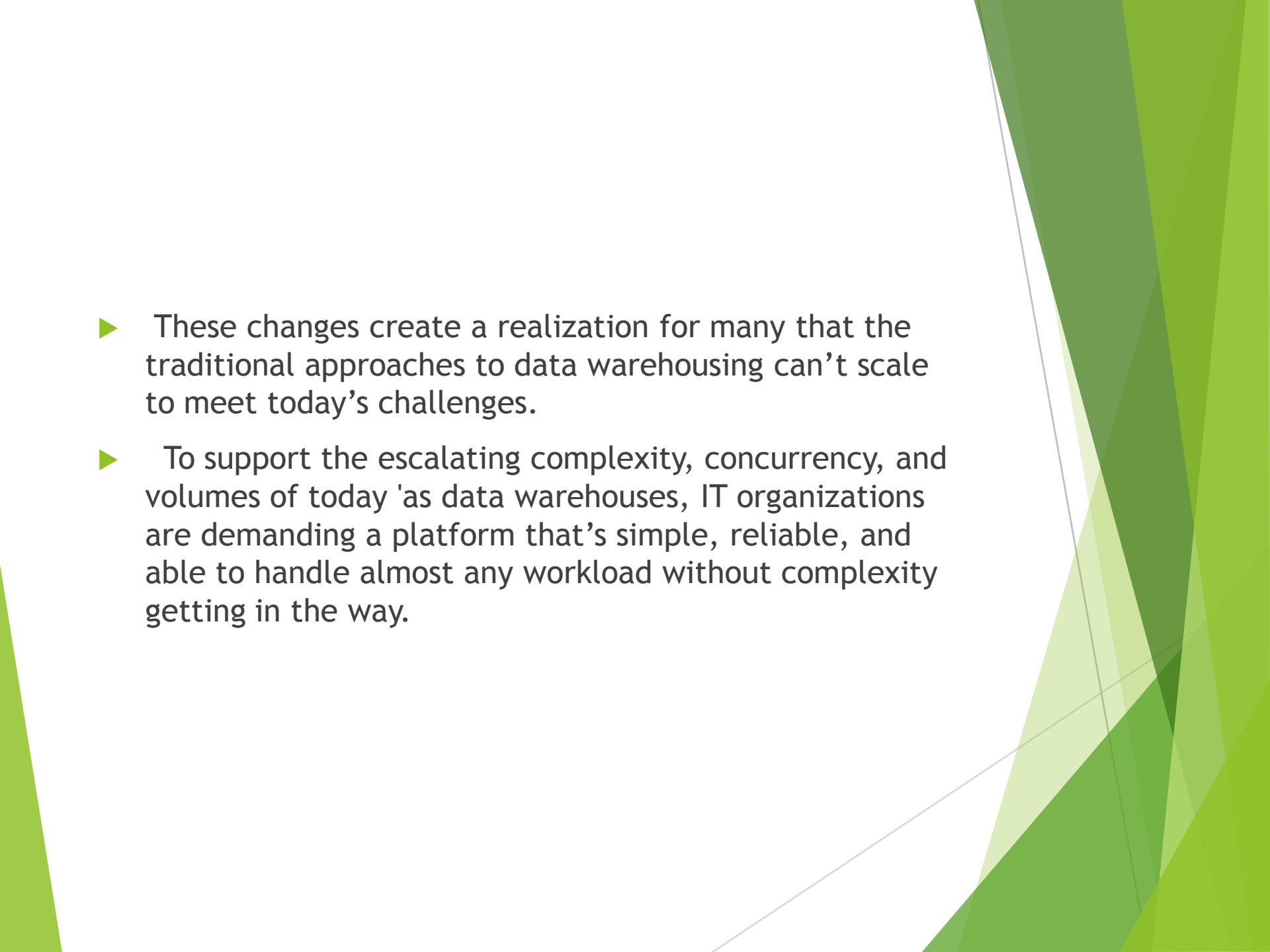
Challenges

- ▶ Financial services companies must bring together a large volume of data, create advanced risk models, and do this quickly without adversely affecting other projects.

Analytics for Big Data at Rest

A Big Data Platform for High-Performance Deep Analytics: IBM Pure Data Systems

- ▶ Data warehousing systems and technologies are intended to give organizations access to information on demand, help them react to it faster, and facilitate quicker decisions.
- ▶ Indeed, the language used to describe the size of data warehouses has changed from gigabytes to terabytes and petabytes, workloads have evolved from being primarily operational to increasingly analytic, and numbers describing concurrency have gone from hundreds to thousands.

- 
- ▶ These changes create a realization for many that the traditional approaches to data warehousing can't scale to meet today's challenges.
 - ▶ To support the escalating complexity, concurrency, and volumes of today's data warehouses, IT organizations are demanding a platform that's simple, reliable, and able to handle almost any workload without complexity getting in the way.

Big Data Platform and Application Frameworks



The IBM Big Data Platform



IBM Blue Gene/O
High performance for
computationally intensive applications



Platform Computing
High performance framework for
distributed computing



InfoSphere Insights
Advanced business intelligence
analytics for existing architectures



InfoSphere Streams
Real-time analytics for
streaming data



**IBM PureScale System for
Operational Analytics**
IBM AS/400 Analytics/DB/OLAP
Data



**IBM Smart Analytics
System**
Operational Analytics on
Structured Data



- ▶ The first generation of data warehouse technologies was modeled after OLTP-based databases running on large symmetric multiprocessing (SMP) machines.
- ▶ These machines had inherent architectural limitations that prevented them from becoming a viable platform for analytics. Subsequent iterations tried to incorporate parallel processing techniques and distributed storage subsystems into the architecture.
- ▶ The resulting complexity of operation (various kinds of indexes, indexes on indexes, optimization hints, and the like) made these systems even more complex to operate and expensive to maintain.

- ▶ Achieving consistent performance against increasing data volumes and diverse workloads without a significant increase in total cost of ownership (TCO) has always been the biggest challenge in data warehousing technologies.
- ▶ Invariably, the biggest bottleneck across all data warehouse operations was the speed at which the database engine could read from and write data to disk, known as disk I/O bottleneck.

IBM PureData System for Analytics Overview

Chris Jackson
Technical Sales Specialist
chris.jackson@us.ibm.com



- ▶ When warehouses were piece-parts, various providers delivered a number of I/O innovations in an attempt to address this bottleneck; however, these innovations were brought to market independently and exhibited little synergy across warehouse tiers: the relational database management system (RDBMS), storage subsystem, and server technologies.
- ▶ The IBM Pure Data System for Analytics appliance—formerly known as the IBM Netezza Data Warehouse Appliance, often just referred to as Netezza— was developed to overcome these specific challenges.

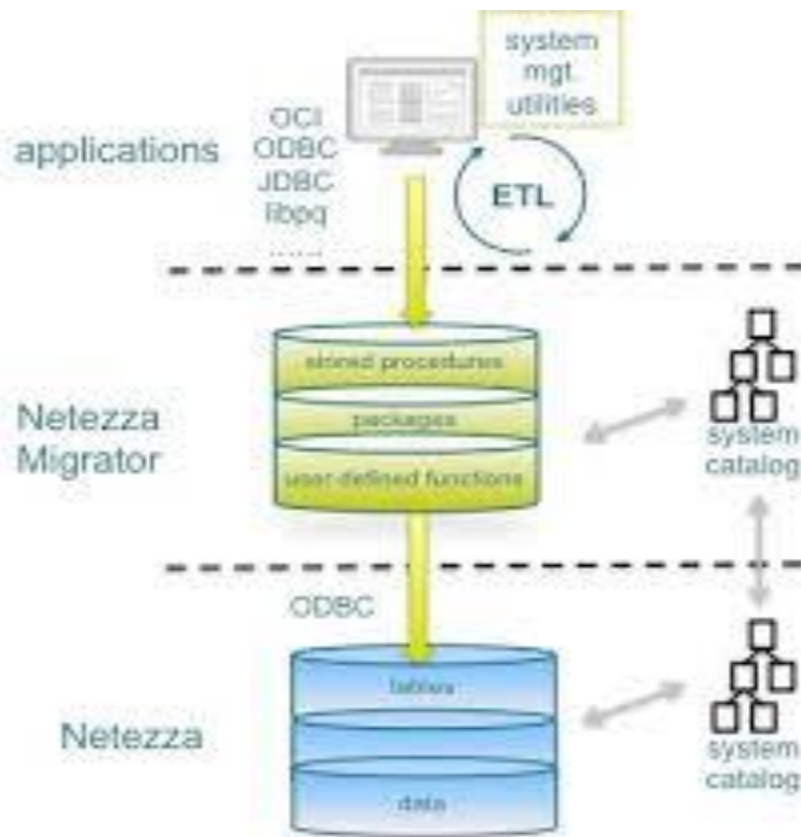
- ▶ It's important to note that Netezza didn't take an old system with known shortcomings and balance it with a storage tier. It was built from the ground up, specifically for running complex analytics on large volumes of structured data.
- ▶ As an easy-to-use appliance, the system delivers its phenomenal results out of the box, with no indexing or tuning required. Appliance simplicity extends to application development, enabling rapid innovation and the ability to bring high-performance analytics to the widest range of users and processes.
- ▶ For users and their organizations, it means the best intelligence to all who seek it, even as demands escalate from all directions.

- ▶ Looking back at the history of Netezza and today's Big Data era, we think it's safe to assert that Netezza took the CFO/CIO discussion from “spending money to save money,” to “spending money to make money”: it handed businesses, both large and small, the ability to democratize deep analytics, while flattening the cost curve that was once associated with this domain.

A Big Data Platform for High-Performance Deep Analytics: IBM Pure Data Systems

Netezza's Design Principles:

- ▶ Netezza's approach to data analysis is patented and proven. Its goal has always been to minimize data movement, while processing it at “physics speed,” in parallel, and on a massive scale—all delivered within an easy-to use appliance at a low cost.



Appliance Simplicity: Minimize the Human Effort

- ▶ Enterprises are spending more and more money to pay people to manage their systems. Now consider this observation in an environment where cheaper labor is presumably available through a globally distributed delivery model, and you can see that the required amount of human effort is a problem.

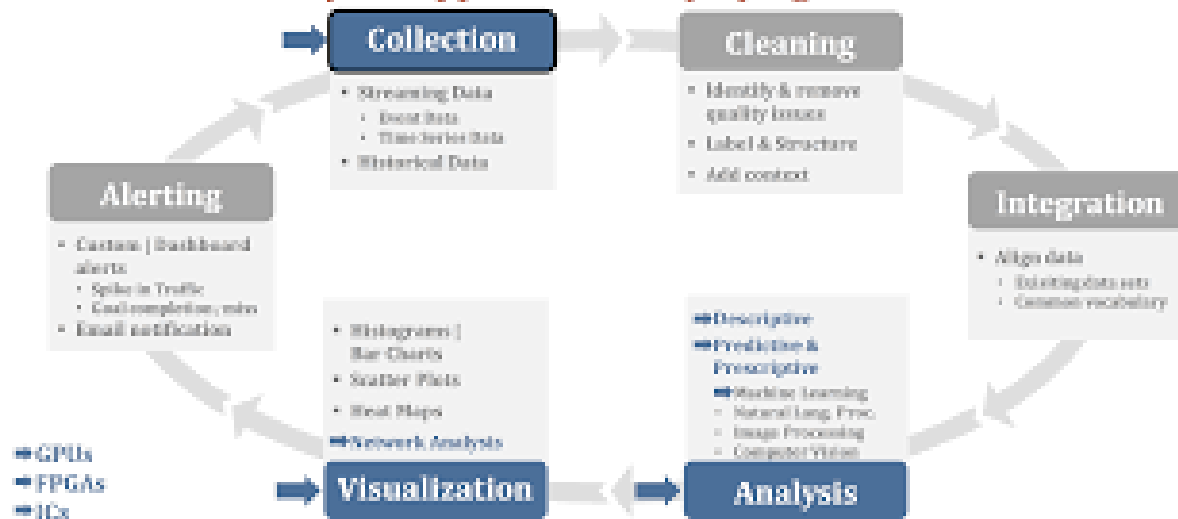


- ▶ Netezza pioneered the concept of appliances in the data warehousing and analytics realm.
- ▶ All of its technologies are delivered in an appliance form, shielding end users from the underlying complexity of the platform.
- ▶ Simplicity rules whenever there is a design tradeoff with any other aspect of the appliance.
- ▶ Unlike other solutions, the appliance just runs, handling demanding queries and mixed workloads at blistering speeds.
- ▶ Even normally time consuming tasks, such as installation, upgrade, and ensuring high-availability and business continuity, are vastly simplified, saving precious time and resources, and mitigates operational risk.

Hardware Acceleration: Process Analytics Close to the Data Store

- ▶ Netezza's architecture is based on a fundamental principle of computer science: when operating on large data sets, don't move data unless you absolutely have to.
- ▶ Moving large data sets from physical storage units to compute nodes increases latency and affects performance. Netezza minimizes data movement by using innovative hardware acceleration;

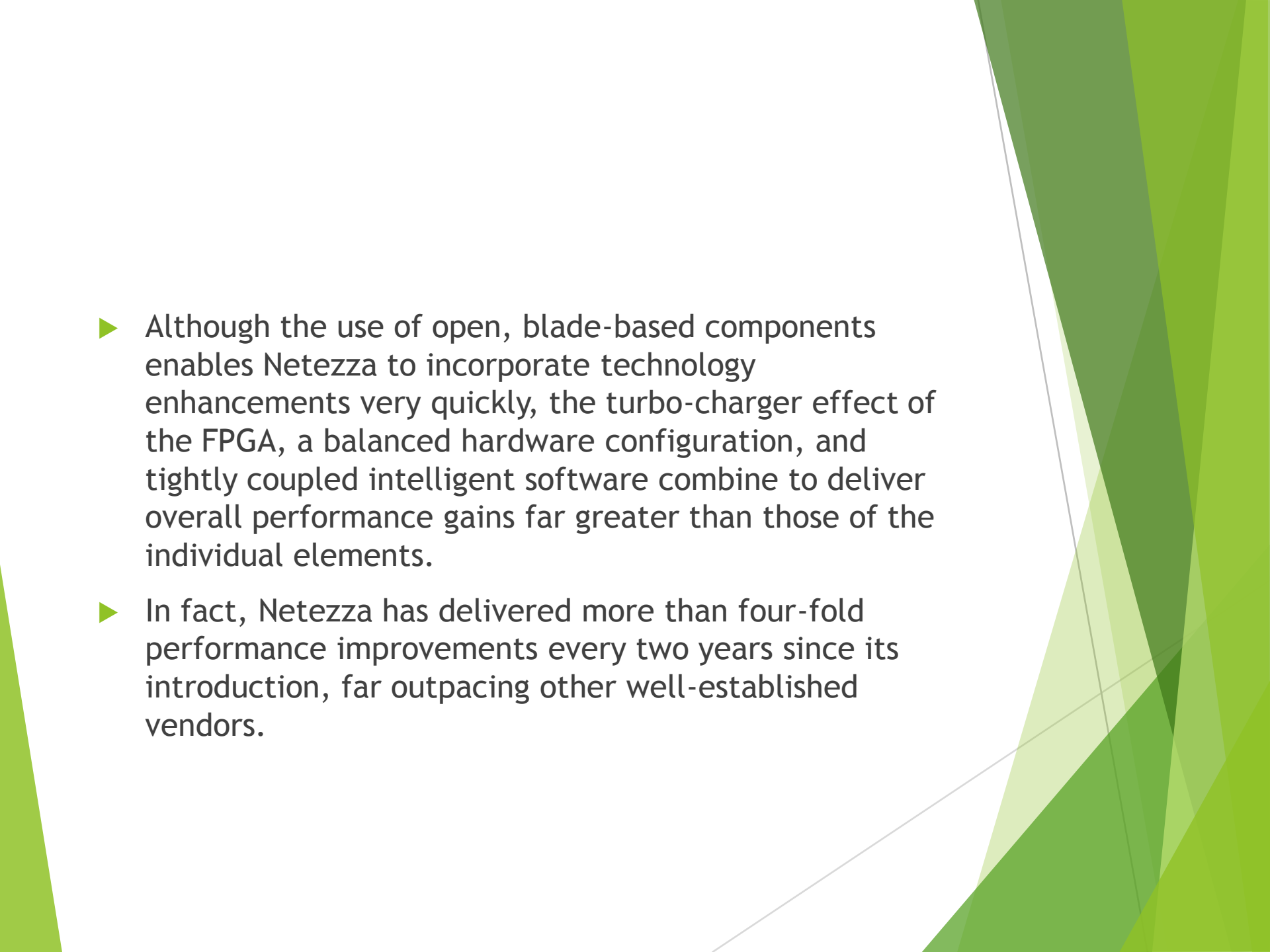
Analytics Applications Employing Hardware



- ▶ for example, it uses field-programmable gate arrays (FPGA) to filter out extraneous data as early in the data stream as possible, and basically as fast as data can be streamed off the disk.
- ▶ This process of data elimination close to the data source removes I/O bottlenecks and keeps downstream components, such as the CPU, memory, and network, from having to process superfluous data; this produces a significant multiplier effect on system performance.

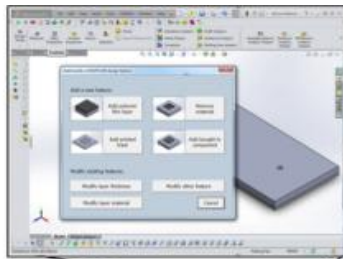
Balanced, Massively Parallel Architecture: Deliver Linear Scalability

- ▶ Every component of Netezza's architecture, including the processor, FPGA, memory, and network, is carefully selected and optimized to service data as fast as the physics of the disk allows, while minimizing cost and power consumption.
- ▶ The Netezza software orchestrates these components to operate concurrently on the data stream in a pipeline fashion, thus maximizing utilization and extracting the utmost throughput from each MPP node.

- 
- ▶ Although the use of open, blade-based components enables Netezza to incorporate technology enhancements very quickly, the turbo-charger effect of the FPGA, a balanced hardware configuration, and tightly coupled intelligent software combine to deliver overall performance gains far greater than those of the individual elements.
 - ▶ In fact, Netezza has delivered more than four-fold performance improvements every two years since its introduction, far outpacing other well-established vendors.

Modular Design: Support Flexible Configurations and Extreme Scalability

- ▶ One of the key concerns with traditional appliance-based architectures has been their ability to scale after data volumes outgrow the physical capacity of the appliance.
- ▶ Netezza addresses this concern with a modular appliance design that simply scales from a few hundred gigabytes to tens of petabytes of user data that can be queried.

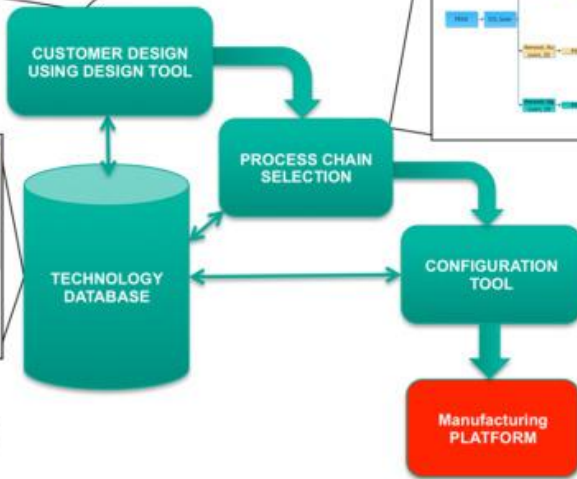


SolidWorks CAD design guidelines

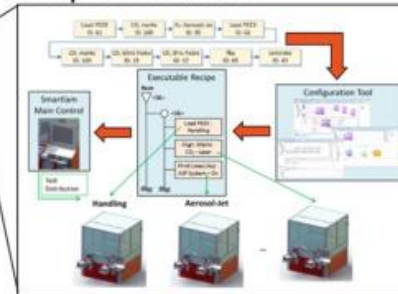
Matlab based numerical process chain selection



Process & material capability repository



Manufacturing control setup & simulation



- ▶ the system has been designed to be highly adaptable, and to serve the needs of different segments of the data warehouse and analytics market.
- ▶ The use of open blade-based components enables the disk-processor-memory ratio to be easily modified in configurations that cater to performance or storage-centric requirements.
- ▶ This same architecture also supports memory-based systems that provide extremely fast, real-time analytics for mission-critical applications.



THANK YOU

**This content is taken from the text books and
reference books prescribed in the syllabus.**

