IoT Academy

Digital Transformation: Enhancing IoT-driven Solutions for Smart Islands

Big data solutions and strategies such as open data and data analytics services for smart islands

Mohammed Al-Ajmi,

Senior Lecturer.
University of Technology and Applied Science, HCT
Mohammed.alajmi@hct.edu.om

Mohammed Khamis Al-Ajmi

- Information Technology Authority
- Head of Quality & Analyst in Digital Forensics - National Digital ForensicLab - Oman CERT
- Analyst and Quality Executive in Digital Forensics - National DigitalForensic Lab - Oman CERT
- Head of Educational Technology, Middle East College
- Author of Books (Arabic):
  - Guide To Microsoft Servers
  - The 7 Element of Digital Citizenship
- LinkedIn

Course Goal

Train participant about Big Data Solutions, Technologies and strategies that can help them to solve real world problems.

Upon completion of this course, the Participant will be able to:

- Differentiate between Traditional Data solutions and Big Data Solution.
- Understand the Modules of Big Data.
- Explore the Different Big Data Technology options available and learn when to use each option.
- Implement this learnings in real world use cases.

Course Structure

- Traditional Data Solutions vs Big Data Solution
- Architecture Template
- Modules
  - What to architect
  - Best Practices
- Technology Options
  - Define
  - Use Cases
- Real Life Use Cases.

Characteristics of Traditional Data

Numbers, Simple Text

- Generated by applications like Finance, Sales, Payroll
- Well defined schema
- Pre-defined linking
- The data attributes hardly change
- Reside within an enterprise (no cloud)
- Centralized data repository (central server)
- Offline backups (cd, tape backup)

Processing Traditional Data

- Small “distances” between source and sink –instantaneous transfers
  - UI to Database
  - Database to Data Processor to Database
  - Database to Reporting
- Data moves from source to the application code for processing
- Data validation happened at the source (no incomplete /dirty data)
- Use RDBMS to manage data (number, text)
- Pre-summarized and computed data
- Reporting is primarily pre-defined format (fixed reports)

Traditional Solutions Architectures

- Have Single Centralized Data Store
- 3-Tier architectures
  - Presentation Layer (UI)
  - Business Layer (Backend)
  - Data Layer
- Use an App codded either at Home or Brought from the market
- Use Integration through custom Interfaces
- Any Changes require full life-cycle projects

Challenges in Traditional Data Solutions

- Cannot handle Text Processing economically
- Cannot handle Incomplete and dirty data
- High costs for storing text data (H/W, S/W)
- Backups restoration is time-consuming
- High management / licensing costs associated with data Management.
- Any Schema changes take significant time

Big Data Solutions

What is Big Data?

- Gartner: Big Data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.
- Variety (Text, Video, Audio, machine data)
- Volume (Tera and Peta Bytes)
- Velocity (speed of data not under control)
- Veracity (dirty, incomplete, inaccurate)

What Lead / Trigger to Big Data

- Cloud adaption
- Social Media
- Mobile explosion
- Machine generated Data (like IOT sensors)
- Data driven management (need to Analyze data to take decision)

What is defined as Big Data Applications

- Data in Tera or Peta Bytes
- More than one source / form
- Text or media data
- Huge processing loads – required more processors
- Real time stream processing
- Advanced Analytics required may use ML / AI
- Big Deployment on multiple servers.
- Changing user requirements
- Deploying Big Data is Relatively cheaper

Characteristics of Big Data Products

- Open Source (free to use)
- Open Integration technologies / APIs
- High interoperability
- Constantly evolving with new features.
- Immature

Big Data Trends

- Numerous companies / projects focused on Big Data technologies
- Mainly open source
- Cloud focused
- Focus on “one thing” with open interfaces for integrations
- Growth in adoption by startup culture
- Number of immature alternatives in each segment

Software Product Organizations

- New domains are driving new product features
  - Cloud
  - Social Media
  - Mobile
- Big Data considered necessary for cost savings
- Flexible ad-hoc analysis capabilities demand flexible schema
- Advanced Analytics being added to reporting solutions

Enterprise with IT - Banks

- Curious and scared at the same time
- Mandated to look at Social / Cloud / Mobile data
- Competitive pressure to be data driven
- Wait and watch until technology is mature
- Starting off proof-of-concepts
- Moving towards the cloud for cost savings

What is a Big Data Solution Architecture?

- Acquire and assemble “big data”
- Various formats from diverse sources
- Process and persist in scalable and flexible data stores
- Provide flexible open APIs for querying
- Provide advanced analytics capabilities
- Use Big Data Technologies to “knit” the solution than building ground up.

Traditional Data Solution vs Big Data Solutions

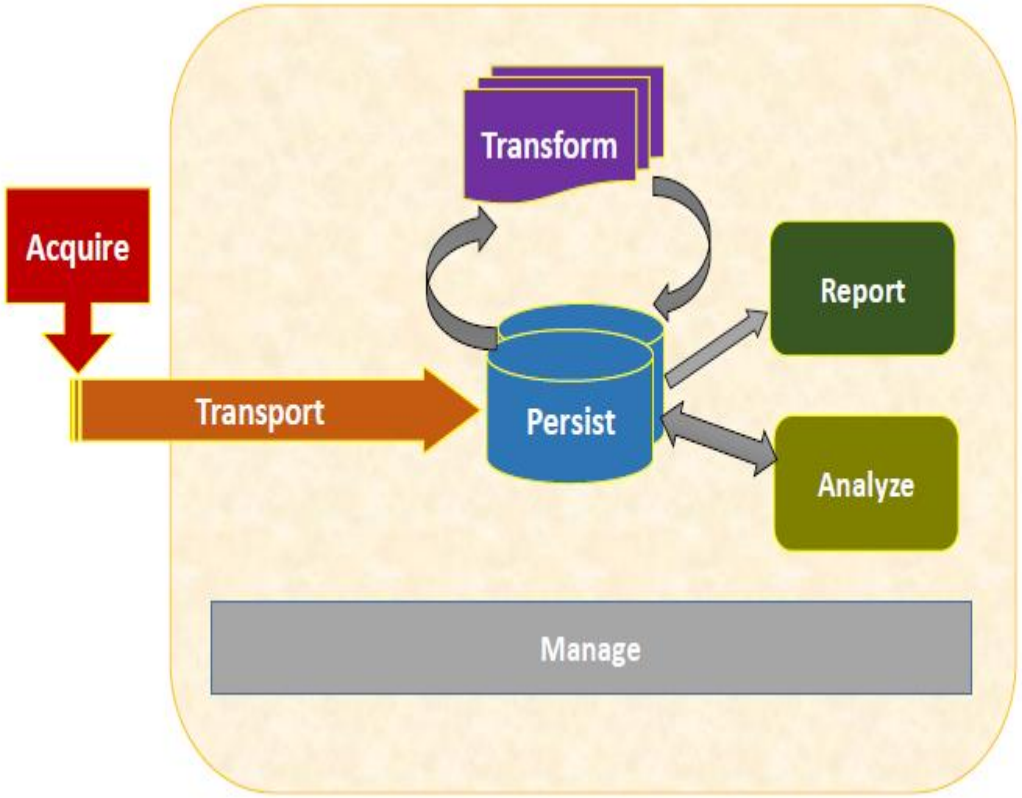
Activity	Traditional Applications	Big Data Applications
Data Acquisition	Data Entry by end users	Fetch ffrom Databases (Traditionnal Applications) , Machine Logs, Social Media
Validation	Validated during Entry	Done post-acquisition – got durty data
Cleansing	Not required	Required for web / social media data
Transformation	Summarization	Text-to-numbers, enrichment, summarization
Persistence	Centralized RDBMS	Distributed Database, Different types of Databases like NOSQL
Applications	3-Tier –business layer centered	Data centered, integration oriented
Usage	Reporting, Analytics,	Analytics, Machine Learning, Predictive & Prescriptive

Historical Data vs Real Time data

Historical Data	Real Time Data
Store-and-forward – pulled Data	Streaming - Pushed Data
End-of-day or end-of-processing trigger(batch processing)	Event based trigger –as it happens
Completed records	Live records with updates
Publish full data/ republish	Can publish part of data
No loss of data , but slow	Possible loss of data , but fast
Detailed analytics	Snapshot / intraday / immediate analytics
Model building	Used to make prediction

Modules of a Big Data Solution (Architecture Template)

- **Acquisition (connect and get Data)**
  - Batch & Stream, multiple formats
- **Transportation (destination BT Source and Sink)**
  - Over internet and organizational boundaries
- **Persistence (store data in different format /places)**
  - Polyglot
- **Transformation**
  - Cleansing, Linking, translating, summarizing
- **Reporting**
  - UIs and APIs
- **Advanced analytics**
  - Machine Learning, Prescriptive
- **Management**



Technologies used for Traditional Data – Build from scratch

- Use an App codded either at Home or Brought from the market
- Single programming language –1000s of LOC written
- Single data store
- High development / maintenance costs

Technologies used for Big Data – Assemble and Stitch

- Big data processing has 2 common demands –scalability and reliability
- A number of products / technologies available especially open source
- They support excellent open integrations
- Acquire most suitable components (solution)
- Stitch /integrate them to create final solution
- Minimal custom work
- Very fast to-production times

Challenges with Big Data Technologies

1. **Too many options**
   - Everyone is coming up with a product
   - Each product addresses a narrow specific field
   - There is no one-product fits-all
   - Everyone is trying to expand to cover other use cases
   - Replacement technologies are invented in a fast pace.
2. **Immature and incomplete**
   - High change rate
   - Field support and services are still primitive
   - Need to still address administration and usability
   - Shortage of skilled and experienced personnel
   - Difficult to predict the future

Challenges with Big Data Technologies

3. **Its future is not safe**
   - Technologies going out of vague before the first release of the application
   - Enterprises like their investments to be safe for at least 10 years
   - Companies supporting most technologies are small / startups.
   - Market deployment size not significant except for a few .

What to expect in the next 5-10 years?

- Few products would grow and become the leaders
- Merging of products
- Fewer more mature options
- Stable features

Making investments and future safe

- Look for product and developer support
- Look at cloud options
- Adaptions by leading companies / products
- Open APIs and data formats

Big Data Solution Modules

1. Acquisition Module

Responsibilities

- Connect / maintain connection to the source
- Execute protocol responsibilities (reconnects, handshakes, error handling)
- Data Format conversion (JSON)
- Filtering
- Local caching
- Compression
- Encryption

Big Data Solution Modules

1. Acquisition Module

Data Source Types

- Databases / Data warehouses
- Files (DB, Logs, Audio, Video, etc)
- HTTP/REST
- Data Streams
- Custom applications

What to architect?

- Identifying new data
- Re-acquisition and retransmit
- Data Loss –not missing records
- Buffering at the source
- Security –source provider policies
- Privacy –policies
- Alerting /alarming for issues

Big Data Solution Modules

1. Acquisition Module

Best Practices

- Involve source owners to establish good handshakes
  - Identifying new data
  - Identifying missing data and retransmission
- Go for reliable Open APIs
- Native APIs/Formats should be standardized as early as possible
- Real-time vs historical –consider separate channels
- Pay attention to security and privacy

Big Data Solution Modules

1. Acquisition Module

Acquire Technology options

1. SQL Query

- Traditional way of extracting data from Relational Databases
- Mature technology
- Ability to transform (joins, group by, cube) and filter data
- Indexing takes care of performance without any programmer work.
- Encryption and compression supported.
- **Use Cases:**
  - RDBMS sources.
  - Apache Hive (open source)

Big Data Solution Modules

1. Acquisition Module

Acquire Technology options

2. Files

- Simple and common way of exchanging / moving data
- Data converted to files (CSV, TSV, XML, JSON)
- Media, text files can be easily stored in files
- **Use Cases:**
  - Inter-organizational data movement
  - Media files
  - Secure data encryption/ compression

Big Data Solution Modules

1. Acquisition Module

Acquire Technology options

3. Rest APIs

- Rest API is A web-based API standard for exchanging data and performing CRUD operations
- Decouples consumers from producers
- Provides Stateless existence
- It Uniform interface across sources : GET, POST, PUT, DELETE
- Support advanced security and encryption
- Support by most cloud and mobile data sources (Twitter, Facebook, Salesforce etc.)
- **Use Cases:**
  - Inter-organizational data movement
  - Media files
  - Secure data encryption/ compression

Big Data Solution Modules

1. Acquisition Module

Acquire Technology options

4. Streaming – push data to client

- Real time data subscribe / publish model
- Client subscribes to a specific topic/ sub-set of data
- HTTP connection is kept “open”
- Server pushes data to client whenever new data is available
- Uses secure keys and encryption
- **Use Cases:**
  - Real time sentiment analysis (Analyze social media data)
  - Real time reporting
  - Real time actions based on user behavior

Big Data Solution Modules

2. Transport Module

Transport Types

- Store and Forward
  - Receive from a source “at rest”
  - Move data in units
  - Track completion
  - Retransmit if required
- Streaming
  - Continuous moving data stream
  - Throttle at source
  - Throttle for sink
  - Inflight storage

Big Data Solution Modules

2. Transport Module

Responsibilities

- Maintain link with acquisition module
- Translate data to protocol optimal formats
- Move data
- Secure Data
- Maintain link with Persistence module
- Save data in Persistence module (and confirm)
- Track data as it moves
- Re-transport in case of failures
- Reporting

What to architect?

- Speed
- Throttling
- Reliability of data (no loss in transport)
- Redundancy
- Scalability
- Status Reporting and alarming
- Compression
- Encryption

Big Data Solution Modules

2. Transport Module

Best Practices

- Do not reinvent the wheel
- Piggy back on proven messaging/transfer frameworks and protocols.
- Look for integrations between transport technologies with acquisition, transformation and persistence technologies.
- Be aware of data transport costs
- Use reliable storage
- Consider security measures to prevent data theft.

Big Data Solution Modules

2. Transport Module

Transport Technology options

1. File Move / Copy commands
    - Simplest way of moving large files
    - Supported on all operating systems
    - Inter-operating system transfers would require adapters
    - Can be quickly scheduled /automated
    - **Use cases:**
        - Intra-enterprise.
        - Media Files.

Big Data Solution Modules

2. Transport Module

Transport Technology options

2. Secure File Transfer Protocol (SFTP)
    - Is Network Protocol for File Access and Transfer
    - Uses a secure channel for data protection (SSH)
    - Support Authentication/authorization (SSH)
    - Support Data integrity checks
    - Can resume interrupted transfers
    - Files carry basic source attributes like timestamps
    - Wide support across O/S, tools and utilities
    - **Use cases:**
        - Inter-enterprise file sharing, log files.
        - Media Files transfers.

Big Data Solution Modules

2. Transport Module

Transport Technology options

3. Apache Sqoop (open source)

- A Command Line tool for transferring data between relational databases and Apache Hadoop
- Allow to transport Entire databases, tables or results of SQL
- Support various file format for Avro, Sequence, Parquet or Plain text files
- Can transport data to Hive and HBase Databases.
- Support Parallelism
- Support Incremental transfers
- BLOB support
- **Use cases:**
  - Hadoop based backups and data warehouses
  - Move data to HBase/Hive
  - Analyzed data from Hadoop to RDBMS

Big Data Solution Modules

2. Transport Module

Transport Technology options

4. Apache Flume (open source)
    - Is A distributed Service for collecting, aggregating and moving large amounts of log/ streaming data
        - Each origin has a source component to get events
        - A Channel used to transport data
        - A sink where the event is deposited
    - Sources can span a large number of servers
    - Support for multiple sources (files, strings, HTTP POSTs, twitter streams)
    - Support for multiple sink types
    - Can add custom sources and sinks through code
    - Robust, fault tolerant, has throttling, failover and recovery capabilities
    - Inflight data processing through interceptors

Use Cases:

- Web log shipping
- Twitter streaming
- Edge server events processing

Big Data Solution Modules

2. Transport Module

Transport Technology options

5. Apache Kafka (open source)
    - An open source message broker platform for real time data feeds
    - Has Publish –subscribe architecture.
    - Developed by LinkedIn, written in Scala.
    - Topics are published. Multiple subscribers can be there for a topic
    - Ordering guarantees
    - Coding required for the publisher and the subscriber to interface with Kafka
    - Support Replication and high availability

Use Cases:

- Real time analytics
- Operational metrics aggregation
- Complex event processing

Big Data Solution Modules

3. Persistence Module

Responsibilities

- Offer reliable data storage.
- Comply with (Atomicity, consistency, isolation, durability)
- Can provide schema.
- Support logical Transactions
- Adopt various Data Access (drivers).
- Support strong response times.
- Support scalability (multi-cluster,..)

What to Architect?

- Scalability
- Consistency
- Transactions
- Read intensive vs write intensive.
- Mutable vs immutable data.
- Able to catalog, get meta-data.
- Latency : real time data vs historical data.
- Standard vs adhoc loads.
- Flexible schema.

Big Data Solution Modules

3. Persistence Module

Best Practices

- Keep schema /design flexible
- Keep data at lowest granularity – at transaction level
- Summarize data only if needed
- Consider your real time application needs
- Take backups

Big Data Solution Modules

3. Persistence Module

Persist Technology Option

1. RDBMS

- Still has a Big role in Big Data Architectures
- Stores data in Tables and Columns
- Optimized for numbers
- Excellent Query performance
- Limitations in scalability.
- Schema need to be predefined
- Mature options –Oracle, MySQL, SQL Server

Use Cases:

- Use it When we need Meta Data
- Multi Update cases – live update
- Work In Progress Data
- Store Summary data
- Store Results

Big Data Solution Modules

3. Persistence Module

Persist Technology Option

2. HDFS (open source)

- A distributed file system that can span across hundreds of data nodes
- Multiple copies of the same file eliminates need for backups
- Can run on commodity servers
- Resilient to node failures.
- Open Source Apache project
- Allows for parallel execution of Map Reduce tasks

Use Cases:

- Store and process Log files
- Media files (recordings)
- Online backup for RDBMS data

Big Data Solution Modules

3. Persistence Module

Persist Technology Option

3. Cassandra (open Source)

- Wide Column Big Table data store.
- Open source developed by Facebook
- Dynamic Schema
- Decentralized architecture
- Single index for each table
- Excellent single-row query performance
- Has Bad range scan performance
- No aggregation support

Use Cases:

- Required to provide Customer 360 view of data.
- Monitoring Statistics and analytics
- Location based lookup

Big Data Solution Modules

3. Persistence Module

Persist Technology Option

4. MongoDB (open Source)

- Document Oriented Database (JSON)
- Strong consistency
- Expressive Query Language
- Support Multiple Indexes on table.
- Support different types of Aggregations
- Support Replication and failover
- Based on Master Slave Model

Use Cases:

- Store documents
- Write once read many data stores
- Real time analytics
- Possible RDBMS replacement

Big Data Solution Modules

3. Persistence Module

Persist Technology Option

5. Neo4j (open Source)

- Graph oriented database
- Deals with Relationships –nodes and edges
- Has ACID Compliant similar to RDBMS
- Transaction support
- Cypher –Graph Query Language
- Easy to program complex joins

Use Cases:

- Master Data Management
- IT Network modeling
- Social network modeling
- Identity and Access Management

Big Data Solution Modules

3. Persistence Module

Persist Technology Option

6. ElasticSearch (open Source)

- A full text search-engine
- A distributed document store
- Every field is indexed and searchable
- Can scale hundreds of servers for structured and unstructured data
- Aggregation support

Use Cases:

- Not recommended as primary data store
- Adhoc query building and aggregation
- Real time analytics.

Big Data Solution Modules

4. Transformation Module

Responsibilities

- Cleansing data
- Filtering
  - Unwanted, incomplete
- Standardization
  - Format, content
- Enrichment
  - Adding ID-names mapping, categorization
- Integration
  - Between data sources
- Summarization

What to architect?

- Real time vs historical
- Create Templates to process data.
- Your data DE normalization
- Reprocessing
- Parallelism
- Speed
- Need Work-In-Progress Storage

Big Data Solution Modules

4. Transformation Module

Best Practices

- Keep Real time and historical separate for data > Tera bytes
- Use Map-Reduce concept for parallel processing
- Don’t reinvent the wheel
- Build template code /functions for enterprise known use cases
- Keep intermediate data for some time to enable reprocessing
- Summarize only if really required. Keep data in original detail
- Build monitoring capabilities for performance

Big Data Solution Modules

4. Transformation Module

Transform Technology options.

1. Custom Code

- Write custom code in your favorite programming language
- Build-it-yourself. Think before you got there
  - Scalability
  - Reliability
  - Parallelism
- People who built such actually ended up building products that we use today

Use Cases:

- Not recommended unless your use case has no readymade solutions

Big Data Solution Modules

4. Transformation Module

Transform Technology options.

2. Hadoop Map-Reduce

- The first big data processing technology
- Processing Code moves to where data resides
- Mappers code works in parallel on individual records and transform
- Reducers code summarize and aggregate data
- Series of map-reducers work on a pipeline
- Uses cheap hardware with extreme parallelism

Use Cases:

- Batch mode processing
- Text mining
- Data Cleansing & Filtering
- Analyzing media files

Big Data Solution Modules

4. Transformation Module

Transform Technology options.

3. *QL Query

- Data Products today have some form of SQL support either native or through other interface
  - Hive, Impala
- Can use *QLs to do:
  - Filtering
  - Cleansing
  - Transformation
  - Summarization
  - Insert/update back to source
- The SQL engine does the heavy lifting

Use Cases:

- Filtering
- Summarization
- Copying Data

Big Data Solution Modules

4. Transformation Module

Transform Technology options.

4. Apache Spark (open source)

- New Generation general data processing engine.
- Eliminates a number of issues in traditional Map-Reduce
- Works on data in memory and in distribution fashion
- Supports Map-Reduce type operations, but a lot faster.
- Can work on Streaming real time data
- Support for Java, Python, R
- Has SQL and Graph capabilities
- Interactive processing capabilities

Use Cases:

- Wide range of use cases
- Interactive processing
- Stream processing

Big Data Solution Modules

4. Transformation Module

Transform Technology options.

5. ETL Products

- Talend, Pentaho, Jaspersoft, Snaplogic.
- Commercial and Open Source offerings
- Visual ETL / pipeline builders with almost no coding
- Can build flows from acquisition to transformation to storage
- Custom functions possible
- Have Operational management modules

Use Cases:

- Any use case is supported on paper.
- Please try out the product before jumping in

Big Data Solution Modules

5. Reporting Module

Responsibilities

- Pre-defined Reports
- Do-it-yourself report designer
- Dashboard Designer
- Need API to extract data from the persistence layer (storage)
- Have Authentication and Authorization
- Real time data presentation
- Alerting

What to architect?

- Focus on Response times
- Access reports from mobile and Mobile.
- Personalization
- Advanced Graphical capabilities
- Threshold management
- Integration with other systems
- Search

Big Data Solution Modules

5. Reporting Module

Best Practices

- Pick a tool with easy to use graphics capabilities
- Tool should have integrated with variety of data sources
- Can Aggregate on the fly without compromising on performance
- Use open standards for data access/integrations
- Provide for personalized dashboards
- Design for multiple interfaces –mobile, web, embedded
- Search is a key capability today

Big Data Solution Modules

5. Reporting Module

Reporting Technology options

1. Cloudera Impala

- In-memory distributed query engine for Hadoop
- Interactive Shell for Queries
- Very fast compared to Hive (no Map Reduce)
- Supports Joins, sub-queries, aggregations
- Supports Hadoop and HBase
- Has ODBC Drivers and Thrift APIs

Use Cases:

- Adhoc querying on Hadoop
- Has API interface for file management applications
- Has HBase data access

Big Data Solution Modules

5. Reporting Module

Reporting Technology options

2. Spark SQL

- Provides SQL like programming –easy to use
- Internally implemented as Map-Reduce operations on Spark .
- Very fast and flexible
- Supports aggregations and joins
- Machine Learning integration with Spark ML
- Can be used for interactive and stream programming

Use Cases:

- Programmed Querying of large datasets
- You have Single system for ETL, Analytics and Advanced Analytics
- Real time analytics

Big Data Solution Modules

5. Reporting Module

Reporting Technology options

3. Elastic

- Elastic has product called ElasticSearch which provides an excellent search engine on existing data
- Has another product called Kibana which provides visualization capabilities on ElasticSearch data
- Has Aggregation capabilities
- Has Streaming data support
- Has Graphics support
- Scalability
- Search support.

Use Cases:

- Enterprise Dashboards and Reports
- Adhoc Query UIs
- Real time Monitoring

Big Data Solution Modules

6. Advanced Analytics Module

Types of Analytics

Types of Analytics	Description
Descriptive	Understand what happened
Exploratory	Find out why something is happening
Inferential	Understand a population from a sample
Predictive	Forecast what is going to happen
Causal	What happens to one variable when you change another
Deep	Use of advanced techniques to understand large and multi-source datasets

Big Data Solution Modules

6. Advanced Analytics Module

Responsibilities

- Model building capabilities
  - Supervised and unsupervised
- Support Validation techniques
- Support Ensemble algorithms
- Support Interactive development
- Has Automation capabilities
- Predictions

What to Architect?

- Scalability
- Performance –especially predictions
- Validations –both model and predictions
- Algorithms –options, configurations, tuning
- Automation and operationalization

Big Data Solution Modules

6. Advanced Analytics Module

Best Practices

- Architecture aligned with methodology –work with data scientists
- Plan adhoc model building –make sure they do not affect other processing
- All Advanced Analytics projects don’t yield results
  - Set expectations right
  - Choose easy projects initially
- Keep an eye on automation and operationalization

Big Data Solution Modules

6. Advanced Analytics Module

Advanced Analytics Technology Options

1. R

- Language / Environment for Statistical Computing and Graphics
- Long history of use by Statisticians
- Wide package set for various machine learning algorithms
- Data Cleansing, Transformation and Graphics capabilities
- RStudio –IDE for interactive programming
- Runs on data in memory
- Limited to the memory of the node where it runs

Use Cases:

- Interactive Model building / Trials on small data sets
- Small data science applications
- Presentations

Big Data Solution Modules

6. Advanced Analytics Module

Advanced Analytics Technology Options

2. Python

- Standard Programming language that has Big Data / Data Science related packages and capabilities
  - NumPy, SciPy, Pandas, Scikit-Learn
- Vast array of third party libraries
- Data Cleansing and Graphics capabilities
- IDEs available for interactive programming
- Integration with Apache Spark
- Multi purpose language –can be used for other capabilities too

Use Cases:

- Interactive Model building / Trials on small data sets
- Data Cleansing
- Small Advanced Analytics applications

Big Data Solution Modules

6. Advanced Analytics Module

Advanced Analytics Technology Options

3. Apache Spark

- Apache Spark has ML Library –supports a good set of machine learning algorithms
- ML library uses Data Frames from Spark SQL
- ML algorithms scale horizontally across a cluster
- Can use Java, Scala or Python
- Interactive model building possible –can run on a windows desktop
- Excellent integration with Big Data Sources
- Real time analytics / predictions

Use Cases:

- Predictive modeling for large datasets
- Model building on NoSQL data sources
- Real time predictions

Big Data Solution Modules

6. Advanced Analytics Module

Advanced Analytics Technology Options

4. Commercial Software

- Tableau, SAS, RapidMiner etc.
- Good set of algorithms
- Good graphics support
- Can Scale
- Can use various Data sources
- Very Pricey.

Use Cases

1. Enterprise Data Backup

Use Case Description

- ABC Enterprise currently keeps 18 months of CRM data in RDBMS and 7 years of archived data on tapes.
- ABC Enterprise wants to move from tape backups to HDFS backups
  - Access of data is easier
  - Can use commodity hardware with potential to move to the cloud
  - No offline backups required
- Provide adhoc querying capability on the data

Use Cases

1. Enterprise Data Backup

Characteristics

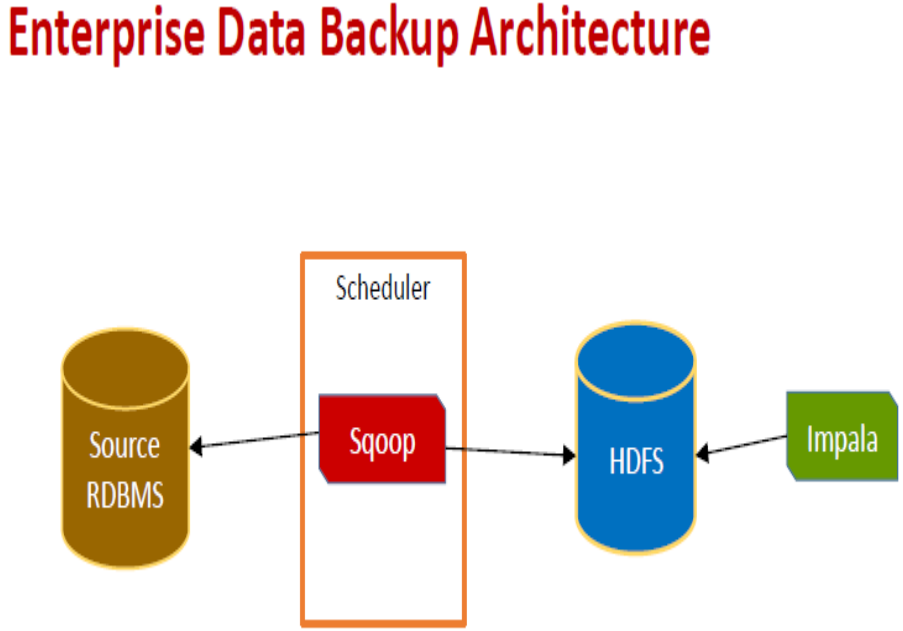
Characteristics	Type	Notes
Sources	RDBMS	
Data Types	Numeric and relational	
Mode	Historical	
Data Acquisition	Pull	
Availability	After 1 day	Data needs to be available in the data warehouse after 1 day since the original data is created
Store type	Write once, read many	
Response Times	As good as possible	Given adhoc querying requirements, queries can run for a few seconds.
Modelbuilding	None	

Use Cases

1. Enterprise Data Backup

Big Data Solution

Module	Technology option	Notes
Acquire	Sqoop	Default choice for Database Extract
Transport	N/A	
Persist	HDFS	Store in native HDFS format as Sequence Files
Transform	N/A	
Reporting	Impala	Basic adhoc querying tool
Advanced Analytic	N/A	



Use Cases

2. Media File Store

Use Case Description

- ABC Enterprise has contact center where all calls are recorded. These recordings need to be archived for analytics
- ABC Enterprise wants to move from tape archive to online archive
- Provide adhoc querying capability on the data

Use Cases

2. Media File Store

Characteristics

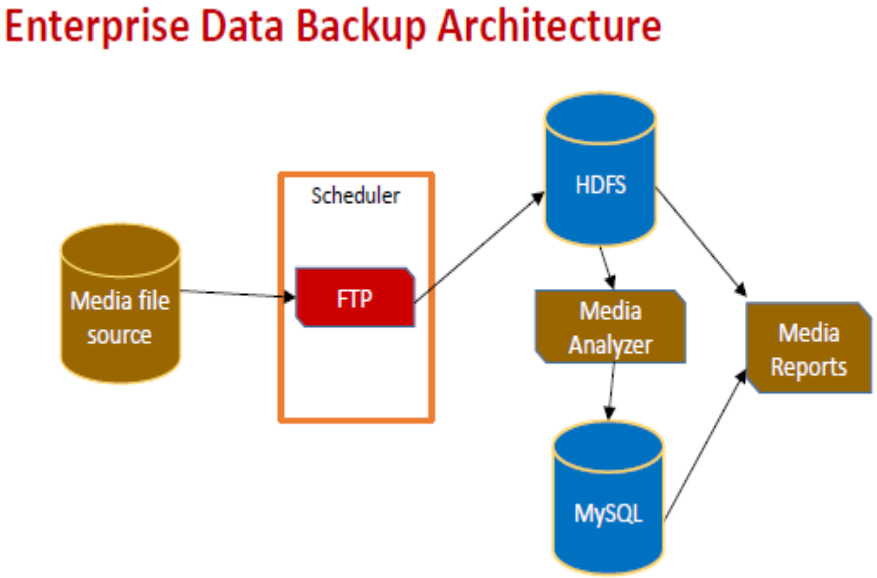
Characteristics	Type	Notes
Sources	Contact Center recording solutions	
Data Types	Media files	
Mode	Historical	
Data Acquisition	Pull	
Availability	After 1 day	Data needs to be available in the media store after 1 day since the original data is created
Store type	Write once, read many	
Response Times	As good as possible	Given adhoc querying requirements, queries can run for a few seconds.
Modelbuilding	None	

Use Cases

2. Media File Store

Big Data Solution

Module	Technology option	Notes
Acquire	Files	Only choice for media files
Transport	FTP	Easy transfer; security and compression capable
Persist	HDFS, MYSQL	Media files stored in HDFS ; Meta-data and analytics stored in MySQL
Transform	Custom	Custom Media Analyzer for tagging media files and storing meta data
Reporting	Impala	Custom Media Reporting tool to analyze meta data and listen to recordings
Advanced Analytic	N/A	



Use Cases

3. Social Media Sentiment Analysis

Use Case Description

- ABC news corporation tracks popular topics on social media and uses them for their news reporting
- They want an automated system to capture social media interactions on popular topics and do real time sentiment analysis
- Sentiment Analysis need to be summarized and archived for future analysis too.

Use Cases

3. Social Media Sentiment Analysis

Characteristics

Characteristics	Type	Notes
Sources	Twitter, Facebook	Social media popular topics. Topics are configurable
Data Types	Tweets, posts (JSON)	
Mode	Real time	
Data Acquisition	Streaming / push	
Availability	Real time	On the fly analytics
Store type	Write many, read many	
Response Times	Real time	Given adhoc querying requirements, queries can run for a few seconds.
Modelbuilding	Sentiment Analysis	

Use Cases

3. Social Media Sentiment Analysis

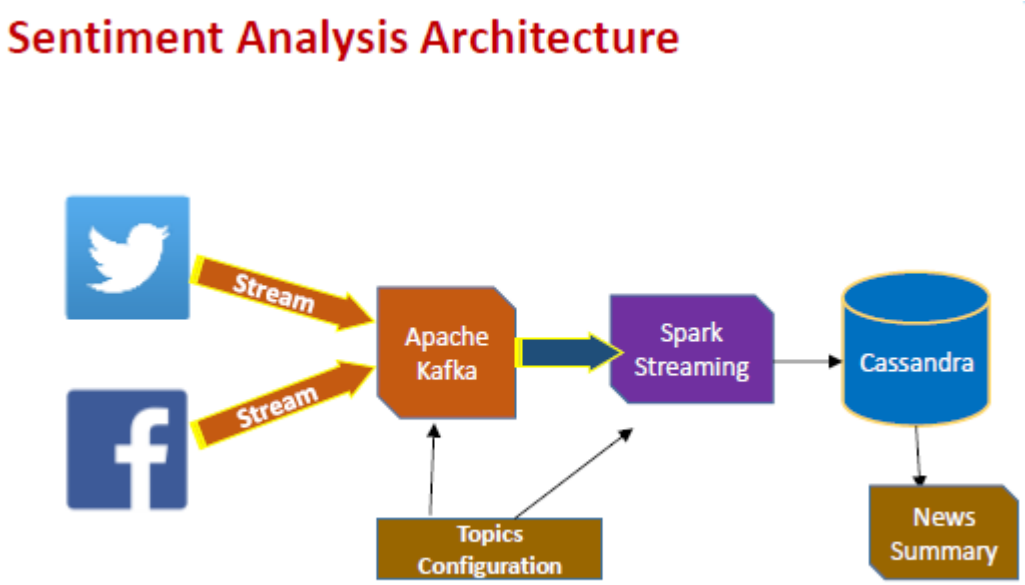
Big Data Solution

Module	Technology option	Notes
Acquire	Streaming	Streaming supported by all social media websites
Transport	Kafka	Kafka provides scalable real time transport for data. Has interfaces to Twitter streaming as well as Spark
Persist	Cassandra	Store data by topic. The social media topic would be used as the key.
Transform	Apache Spark	Real time stream subscription and transformation
Reporting	Custom	Custom application for reading Cassandra data and summarizing for news
Advanced Analytic	Apache Spark	Sentiment Analysis on the fly with stream processing

Use Cases

3. Social Media Sentiment Analysis

Big Data Solution



Use Cases

4. Credit Card Fraud Detection

Use Case Description

- ABC Systems runs a web based retail solution where customers can order any kind of products (like Amazon)
- Sometimes credit card thieves use stolen information to make purchases. This later results in loss of revenue
- ABC systems needs a real time Credit Card Fraud prediction system so that the purchase is blocked before its complete.

Use Cases

4. Credit Card Fraud Detection

Characteristics

Characteristics	Type	Notes
Sources	web transactions	Data is captured in real time while payment is being made on the web
Data Types	Numeric / CRM	
Mode	Real time / Historical	Historical data collection ; prediction in real time
Data Acquisition	Streaming / push	Data pushed from browser as transactions happen
Availability	Real time	Real time predictions
Store type	Write once , read many	
Response Times	Minimal	Prediction need to be made when the purchase is made.
Modelbuilding	Binary Classification	Model to predict if a transaction is fraudulent or not.

Use Cases

4. Credit Card Fraud Detection

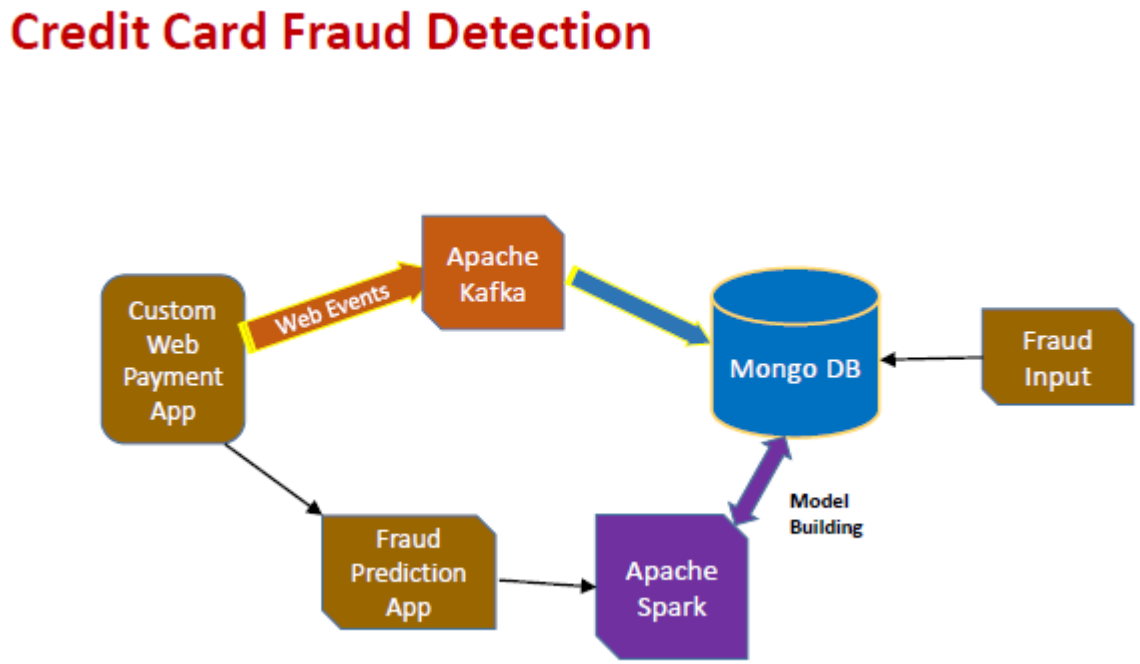
Big Data Solution

Module	Technology option	Notes
Acquire	Web Events	Generated by custom web app. Deployed on a web farm
Transport	Kafka	Kafka provides scalable real time transport for data. Web Transaction events from web app.
Persist	MongoDB	Web events/transactions accumulated and stored in Mongo DB; Also models built are stored in Mongo DB
Transform	Spark	
Reporting	None	Architecture can be enhanced to add adhoc reporting on the web transactions if required.
Advanced Analytic	Apache Spark	Binary Classification model building

Use Cases

4. Credit Card Fraud Detection

Big Data Solution



Use Cases

5. Connected Car - IOT

Use Case Description

- ABC Car company wants to connect cars in real time to analytics engine
- Cars have multiple sensors. Sensor data need to be analyzed (real time / historical) to generate alarms for possible failures to the driver
- ABC needs a satellite enabled data collection and alarm system backed by a big data infrastructure

Use Cases

5. Connected Car - IOT

Characteristics

Characteristics	Type	Notes
Sources	Car sensors	Sensors in car
Data Types	Numbers	Numeric event sensor data
Mode	Historical / Real time	Critical data processed real time. Rest historical
Data Acquisition	Push	Sensors send data to collection centers
Availability	Real time	Real time alarms needed
Store type	Write many, read many	Car profile need to be stored
Response Times	Real time	Real time profile fetches for real time alarming
Modelbuilding	Car issue prediction	Predict possible future issues

Use Cases

5. Connected Car - IOT

Big Data Solution

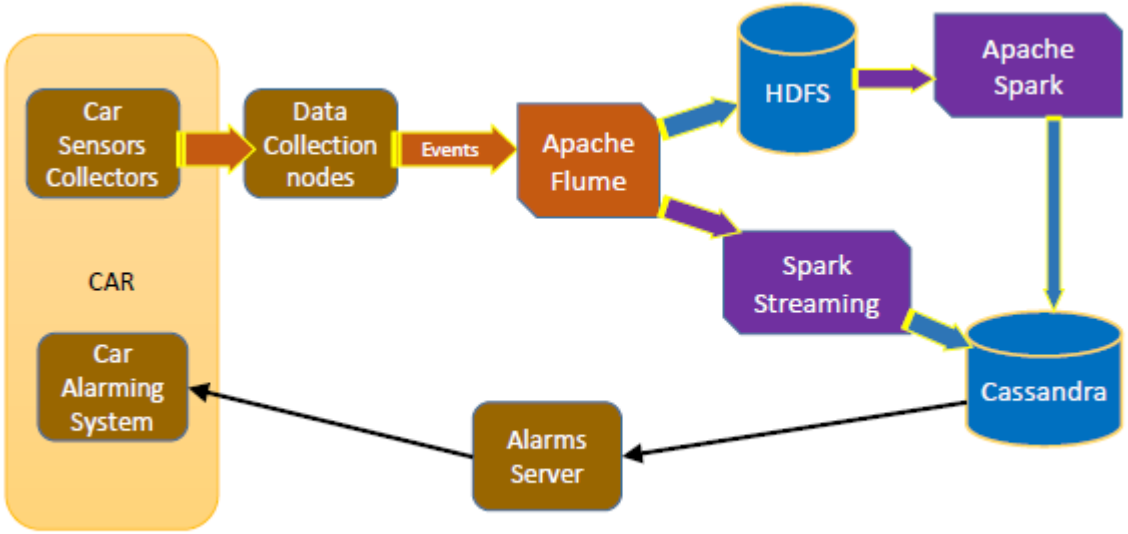
Module	Technology option	Notes
Acquire	Events from Car Sensors	
Transport	?	
Persist	?	
Transform	?	
Reporting	Custom	
Advanced Analytic	?	

Replace the Question Mark (?) with appropriate Technology option for each Big Data Module.

Use Cases

5. Connected Car - IOT

Big Data Solution




thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
tapadh leat
спасибо
dank je
dziękuję
bedankt
hvala
mauruuru
obrigado
sukriya
kop khun krap
go raibh maith agat
arigatô
takk
dakujem
merci
terima kasih
grazie
mochchakkeram
chnorakaloutioun
gracies
gratias ago
sulpay
sagolun
tanemirt
rahmet
dhanyavadagalu
shukriya
merci
мерси
рахмат
Баярлалаа
barka
welalin
tack
mersi
spas
vinaka
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
obrigada
murakoze
tenki
djiere dieuf
tau
дякую
mamnun
trugarez
dankon
akun
kiitos
dankie
dhanyavad
nanni
nandri
bayarlalaa
gracie
enkosi
koszonom
sobodi
dekuji
mesi
didi madloba
najis tuke
kam sah hamnida
rahmat
diolch
xiexie
감사합니다
ευχαριστώ