P01: Big Data Architecture and Patterns

- Objective

- 5-parts:
  1. Big data classification and architecture
  2. How to know if a big data solution is right for your organization
  3. Understanding the architectural layers of a big data solution
  4. Understanding atomic and composite patterns for big data solutions
  5. Apply a solution pattern to your big data problem and choose the products to implement it

- Summary

Objective

1. Defines key concepts for building the architecture for a big data solution

2. Key questions to ask to assess the viability of a big data solution

3. Components of a big data solution from data source to business insight; use of layers to categorized functions

4. Basic patterns to apply to particular situations; from atomic to composite patterns; pattern identifies components required for the solutions

5. Putting it together – from problem to solution

5 Parts*

1. Big data classification and architecture
2. How to know if a big data solution is right for your organization
3. Understanding the architectural layers of a big data solution
4. Understanding atomic and composite patterns for big data solutions
5. Apply a solution pattern to your big data problem and choose the products to implement it

adapted from Big Data Architecture and Patterns, IBM DeveloperWorks, 2013.

1. Introduction to Big Data Classification and Architecture

i. What is the big data type of your problem?

ii. What characteristics do the big data type of your problem have?

Big Data Classification and Architecture

- Definition of big data
  - Volume
  - Velocity
  - Variety

- How to solve big data problem
  - Map big data problem to its big data type
  - Use big data type to classify big data characteristics
  - Determine the appropriate classification pattern
  - Determine the appropriate big data solution

Big Data Classification and Architecture

- Map big data problems to its big data type

- Examples:
  - Telecommunications: Customer churn analytics – Web and social data & Transaction data
  - Marketing: Sentiment analysis – Web and social data

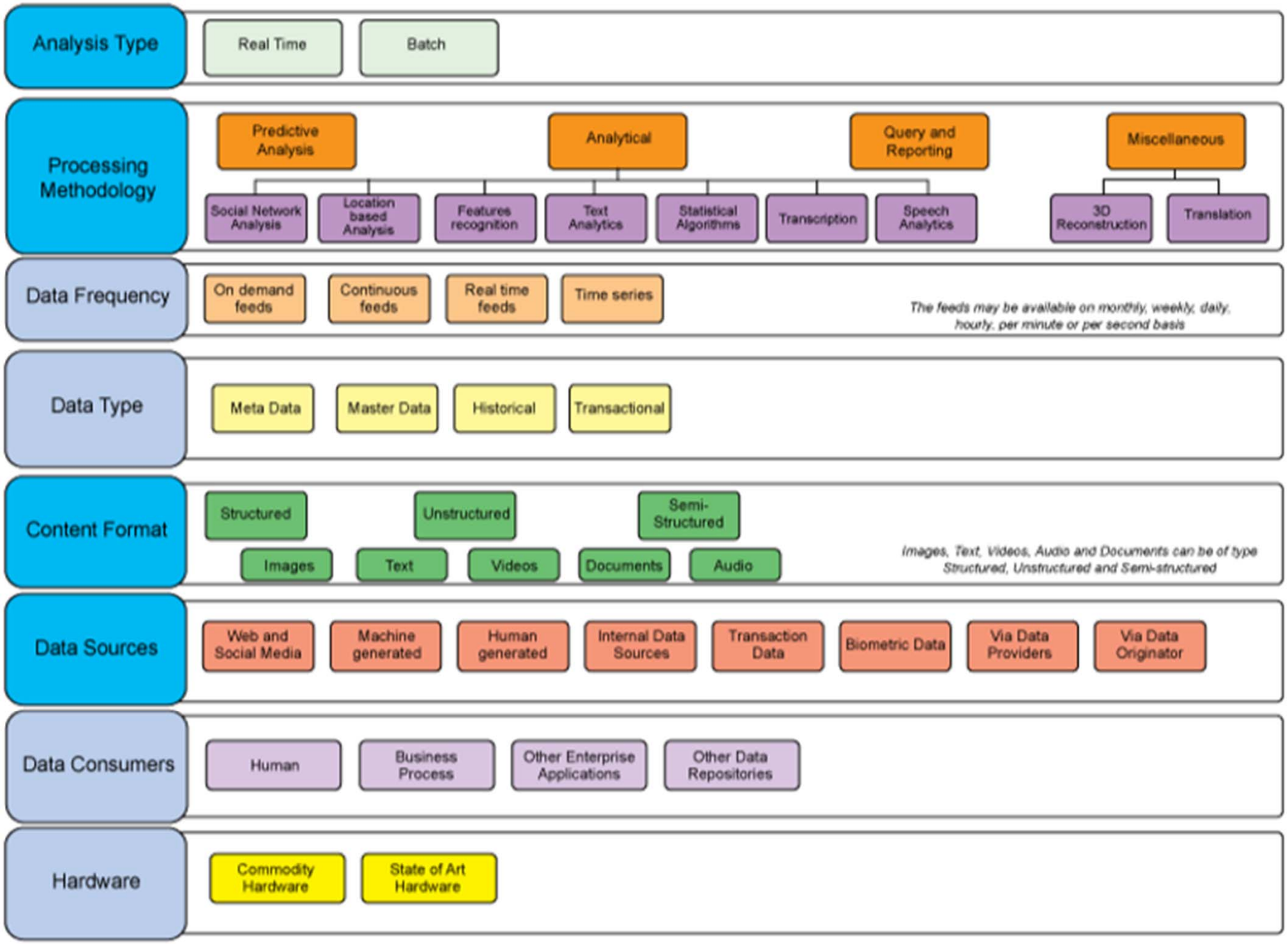
Classifying Big Data Problem according to Type of Big Data

Business problem	Big data type	Description
Telecommunications: Customer churn analytics	Web and social data Transaction data	Telecommunications operators need to build detailed customer churn models that include social media and transaction data, such as CDRs, to keep up with the competition. The value of the churn models depends on the quality of customer attributes (customer master data such as date of birth, gender, location, and income) and the social behavior of customers. Telecommunications providers who implement a predictive analytics strategy can manage and predict churn by analyzing the calling patterns of subscribers.
Marketing: Sentiment analysis	Web and social data	Marketing departments use Twitter feeds to conduct sentiment analysis to determine what users are saying about the company and its products or services, especially after a new product or release is launched. Customer sentiment must be integrated with customer profile data to derive meaningful results. Customer feedback may vary according to customer demographics.

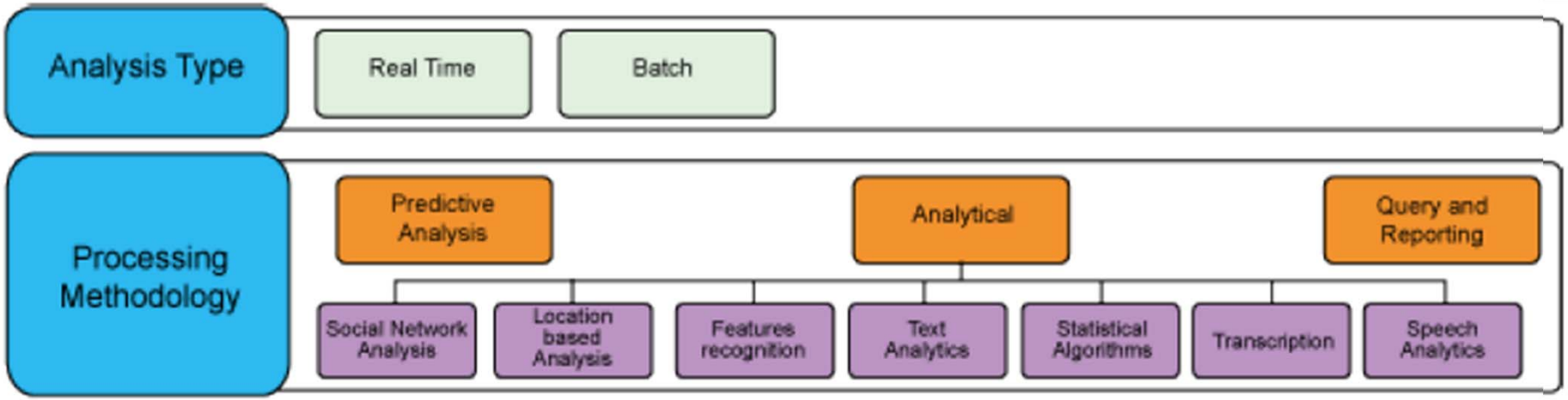
Big Data Characteristics

- Analysis type
- Processing methodology
- Data frequency and size
- Data type
- Content format
- Data source
- Data consumers
- Hardware

Classify Big Data Characteristics



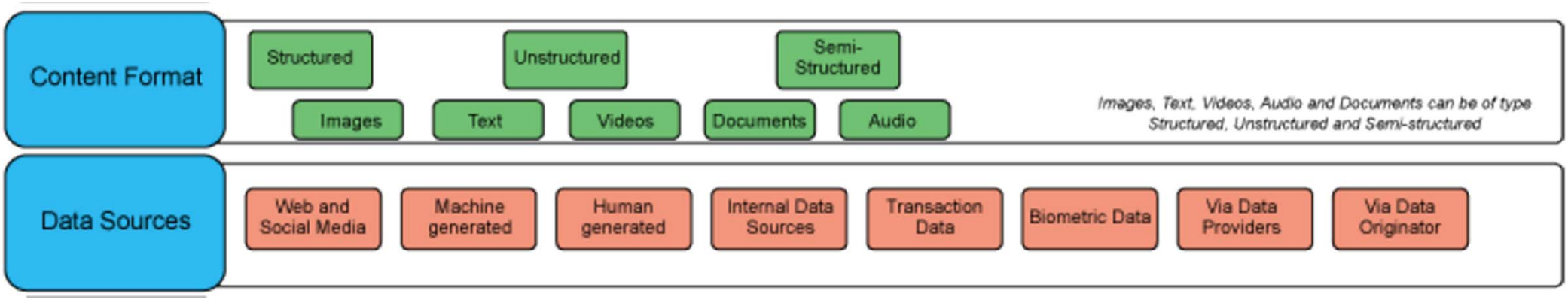
Classify Big Data Characteristics



Analysis type - Whether the data is analyzed in real time or batched, give careful consideration to choosing the analysis type, since it affects several other decisions about products, tools, hardware, data sources, and expected data frequency.

Processing methodology - Techniques to be applied for processing data (e.g., predictive, analytical, ad-hoc query, and reporting). Business requirements determine the appropriate processing methodology. Choice helps identify the appropriate tools and techniques to be used in your big data solution

Classify Big Data Characteristics



Content format: Format of incoming data — structured (RDMBS, for example), unstructured(audio, video, and images, for example), or semi-structured. Format determines how the incoming data is processed and is key to choosing tools and techniques and defining a solution from a business perspective.

Data source - Sources of data (where the data is generated) — web and social media, machine-generated, human-generated, etc. helps determine the scope from a business perspective.

Example: Insurance Fraud

- Insurance fraud is an act or omission intended to gain dishonest or unlawful advantage, either for the party committing the fraud or for other related parties

- **Current fraud-detection process**
  - Monitoring fraud
  - Searching for potential fraud indicators
  - Coordinating with law enforcement agencies

Example: Insurance Fraud

- **Issues with the current fraud-detection process**
  - Limited Data Sets
    - Traditional solutions use models based on historical fraud data, black-listed customers and insurance agents, and regional data about fraud peculiar to a certain area
  - Manual Work
    - Insurers may not be able to investigate all the indicators. Fraud is often detected very late, and it is difficult for the insurer to do adequate follow up for each fraud case
  - Hard to deal with new cases
    - Current fraud detection relies on what is known about existing fraud cases, so every time a new type of fraud occurs, insurance companies have to bear the consequences for the first time

Insurance Fraud's Big Data Characteristics

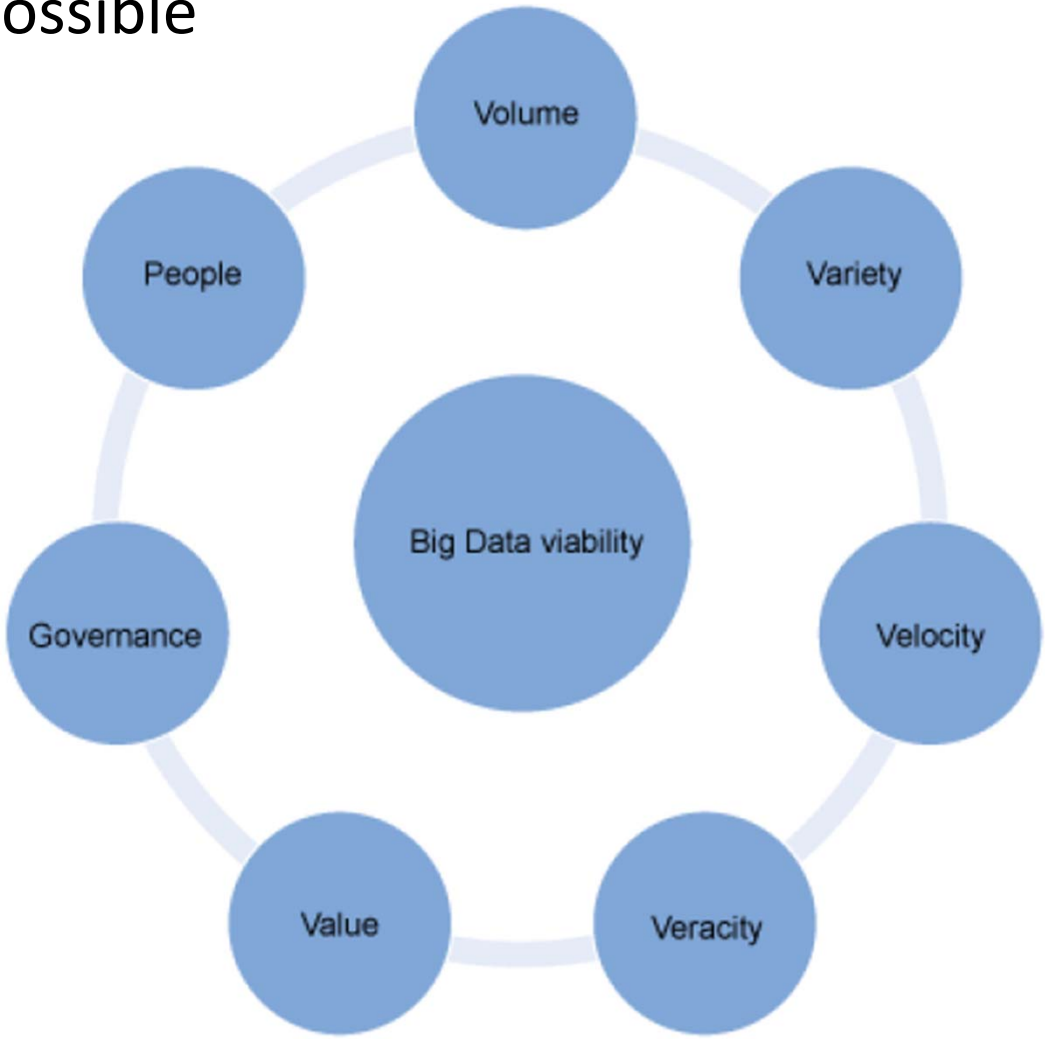
- Analysis Type: Batch
- Processing Methodology: Analytical
- Content Format:
  - Structured(Machine-generated data and transaction data)
  - Unstructured(Human-generated data)
- Data Source:
  - Machine-generated data
  - Transaction data
  - Human-generated data

2. How to Know if a Big Data Solution is Right for Your Organization

- What to consider when accessing the viability of a big data solution?
- How to consider?

Assessing the Viability of a Big Data Solution

- Value or what insights are possible
- Volume of data
- Variety of data sources
- Velocity of data generation
- Veracity of data
- Governance considerations
- People



Value

- Is there existing data that can be used to get insight?
  - organizations have quite a lot of data not being harnessed for business insight, include log files, errors files, and operational data from applications. Don't overlook this data as a potential source of valuable information.

Complexity – Volume

- **Has the volume of data increased?**

- You may want to consider a big data solution if:
  - data is sized in petabytes and exabytes, and in the near future, might grow to zeta bytes
  - data volume is posing technical and economic challenges to store, search, share, analyze, and visualize using traditional methods, such as relational database engines
  - data processing can currently make use of massive parallel processing power on available hardware

Complexity – Variety

- **Has the variety of data increased?**

- variety of data might demand a big data solution if:
  - data content and structure cannot be anticipated or predicted.
  - data format varies, including structured, semi-structured, and unstructured data.
  - data can be generated by users and machines in any format, for example: Microsoft® Word files, Microsoft Excel® spreadsheets, Microsoft PowerPoint presentations, PDF files, social media, web and software logs, email, photos and video footage from cameras, information-sensing mobile devices, aerial sensory technologies, genomics, and medical records.
  - New types of data have emerged from sources that weren't previously mined for insight.
  - Domain entities take on different meanings in different contexts.

Complexity – Velocity

- **Has the velocity of data increased or changed?**

- Consider whether your data:
  - Is changing rapidly and must be responded to immediately
  - Has overwhelmed traditional technologies and methods, which are no longer adequate to handle data coming in real time

Complexity – Veracity

- **Is your data trustworthy?**

- Consider a big data solution if:
  - authenticity or accuracy of the data is unknown
  - data includes ambiguous information
  - unclear whether the data is complete

People

- **Are the right skills on board and the right people aligned?**

- Specific skills are required to understand and analyze the requirements and maintain the big data solution. Skills include industry knowledge, domain expertise, and technical knowledge on big data tools and technologies.

- Before undertaking a new big data project, make sure the right people are on board:
  - Do you have buy-in from stakeholders and other business sponsors who are willing to invest in the project?
  - Are data scientists available who understand the domain, who can look at the massive quantity of data and who can identify ways to generate meaningful and useful insights from the data?

Example: Insurance Fraud

- Insurance fraud is an act or omission intended to gain dishonest or unlawful advantage, either for the party committing the fraud or for other related parties

- In India, the loss in 2011 alone totals INR 300 billion. Apart from the financial loss, insurers are losing business because of customer dissatisfaction.

Example: Insurance Fraud

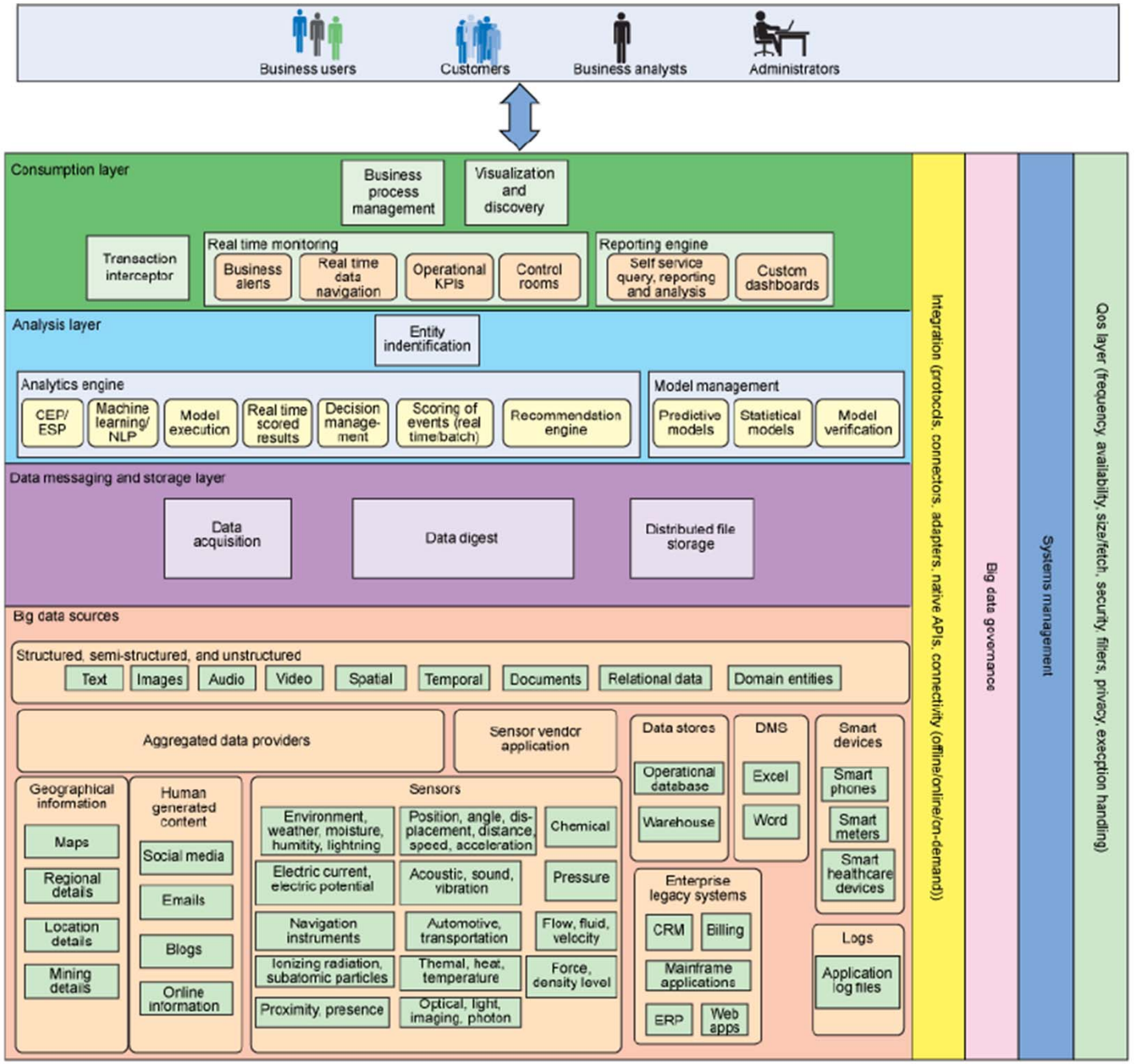
- Value or what insights are possible
  - Reduce loss
- Volume of data
  - Large, and still increasing
- Variety of data sources
  - Transaction/Machine-generated/Human-generated
- Velocity of data generation
  - No need of real time, batch analysis is OK.
- Veracity of data
  - ?
- Governance considerations
  - ?
- People
  - ?

3. Architectural Layers of a Big Data Solution

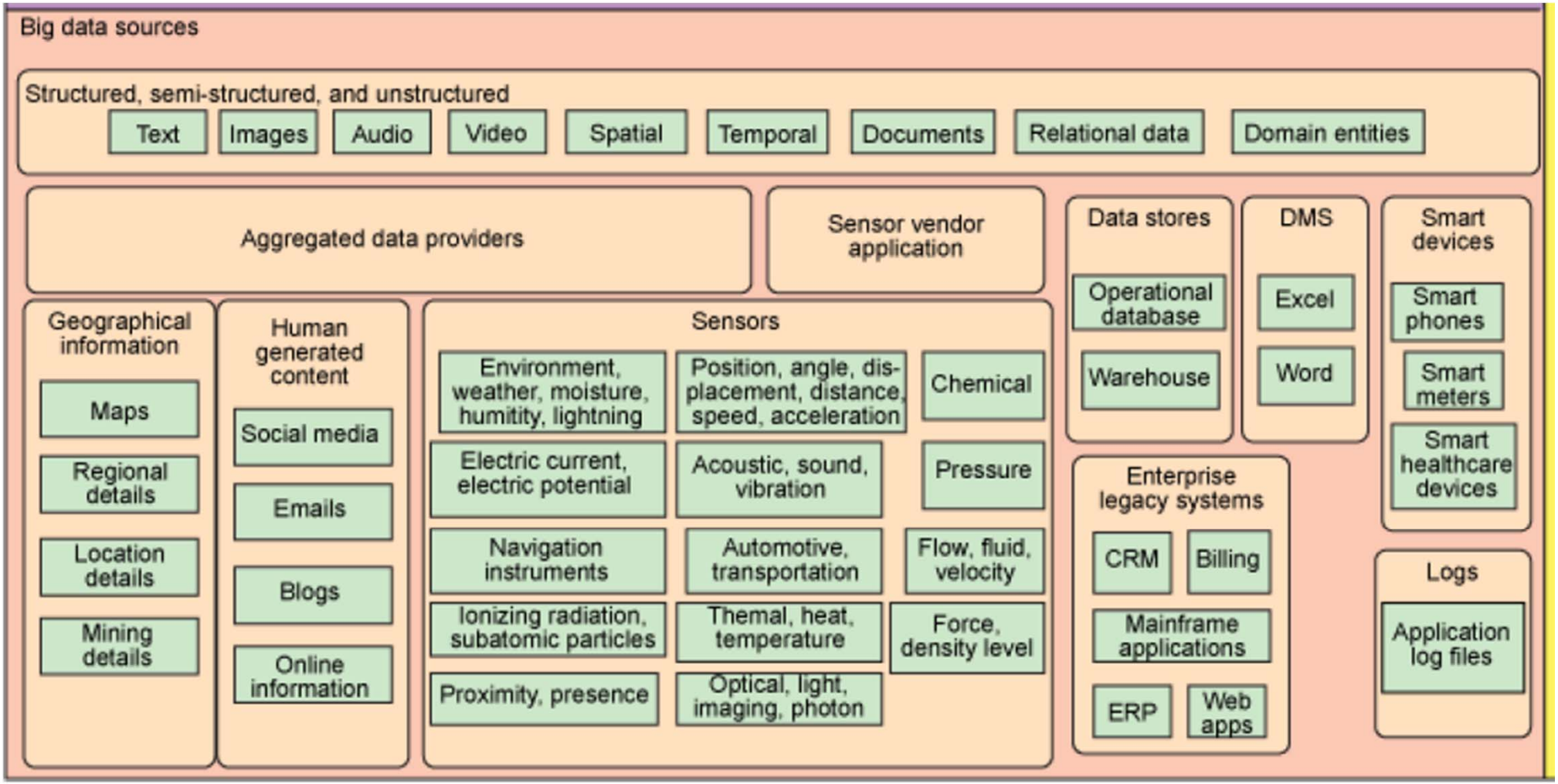
Four Logical layers in a big data solution

i. Big data sources
ii. Data massaging and store layer
iii. Analysis layer
iv. Consumption layer

Logical Layers of a Big Data Solution



i. Big Data Source

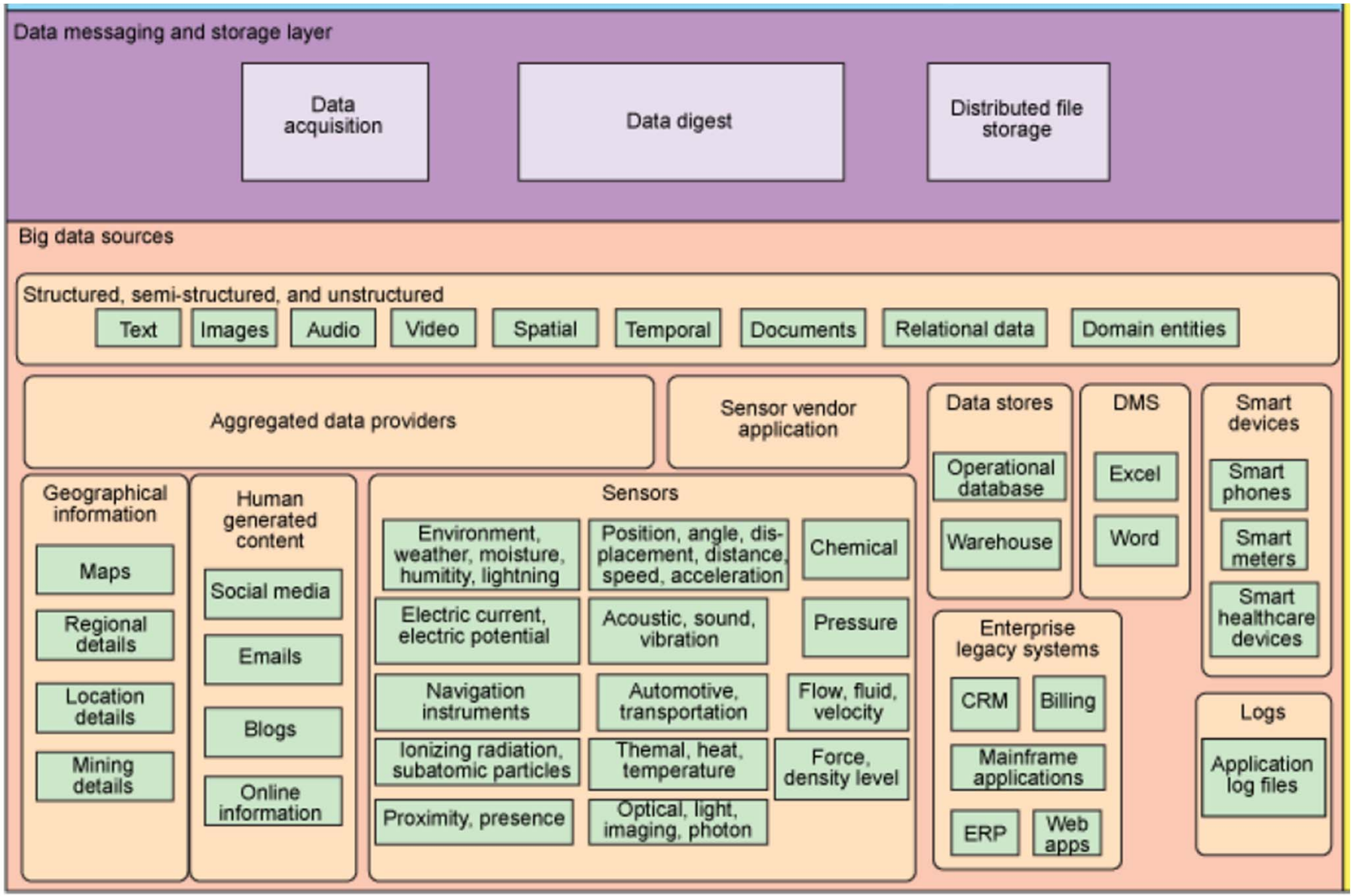


i. Big Data Sources

- data available for analysis comes from all channels

- Data varies in format and origin:
  - **Format** — Structured, semi-structured, or unstructured
  - **Velocity and volume** — speed that data arrives and the rate at which it's delivered varies according to data source
  - **Collection point** — Where the data is collected, directly or through data providers, in real time or in batch mode. The data can come from a primary source, such as weather conditions, or it can come from a secondary source, such as a media-sponsored weather channel
  - **Location of data source** — Data sources can be inside the enterprise or external. Identify the data to which you have limited-access, since access to data affects the scope of data available for analysis

ii. Data Massaging and Store Layer

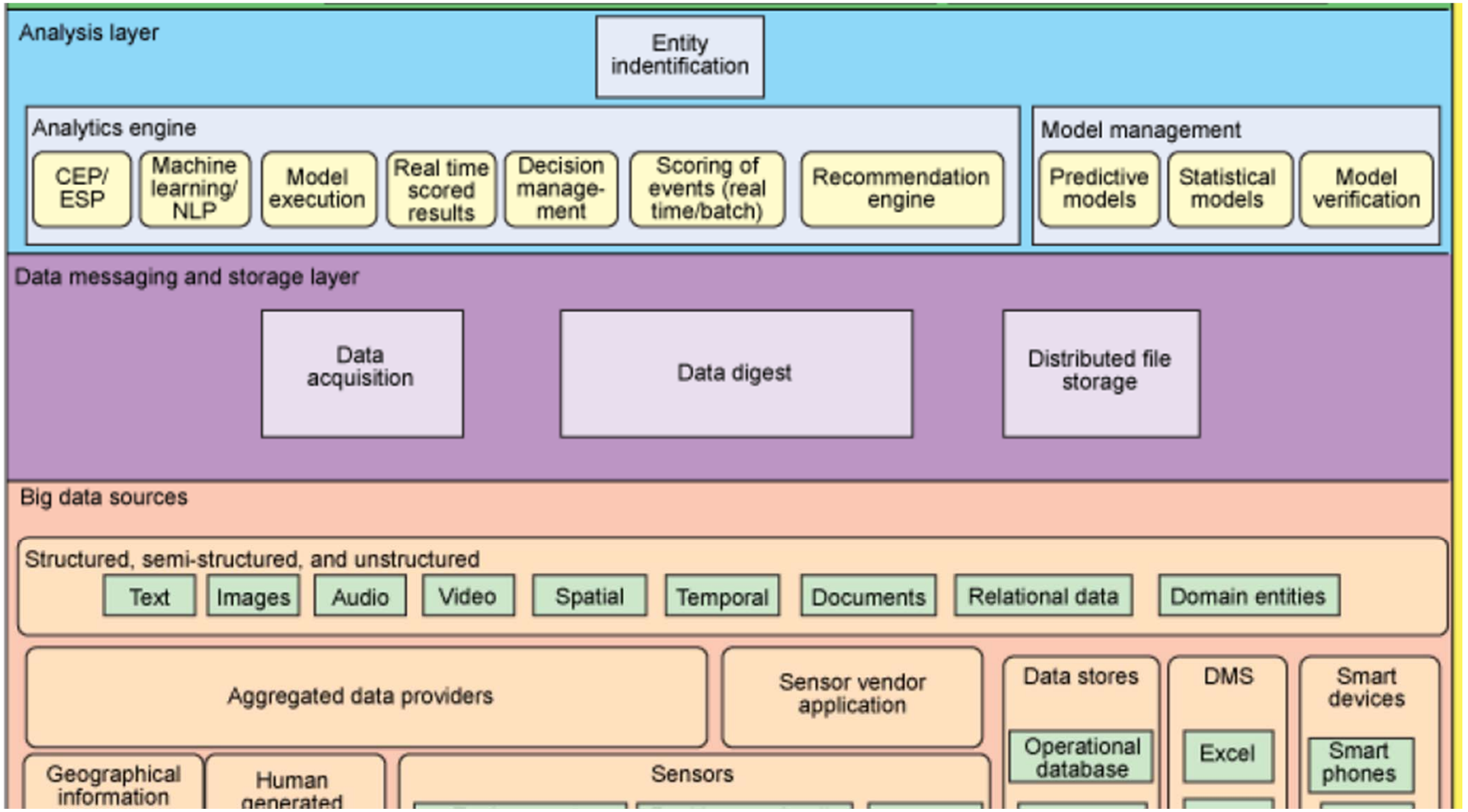


ii. Data Massaging and Store Layer

- responsible for acquiring data from the data sources and converting it to a format for analyzed

- For example, an image might need to be converted so it can be stored in an Hadoop Distributed File System (HDFS) store or a Relational Database Management System(RDBMS) warehouse for further processing. Compliance regulations and governance policies dictate the appropriate storage for different types of data.

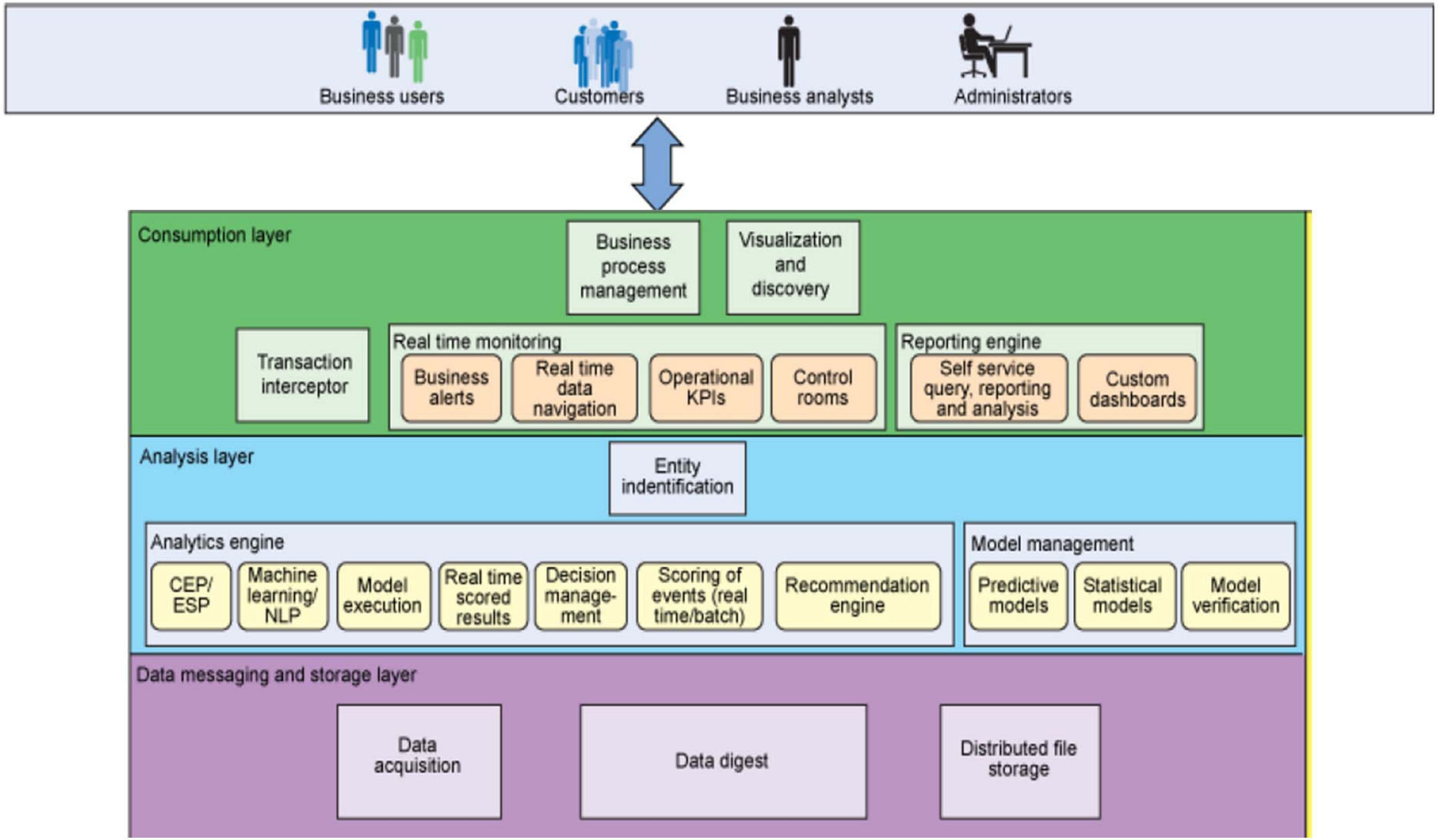
iii. Data Analysis Layer



iii. Analysis Layer

- reads the data digested by the data massaging and store layer
- analysis layer can access data directly from data source
- decisions to be made to manage the tasks:
  - Produce the desired analytics
  - Derive insight from the data
  - Find the entities required
  - Locate the data sources that can provide data for these entities
  - Understand what algorithms and tools are required to perform the analytics.

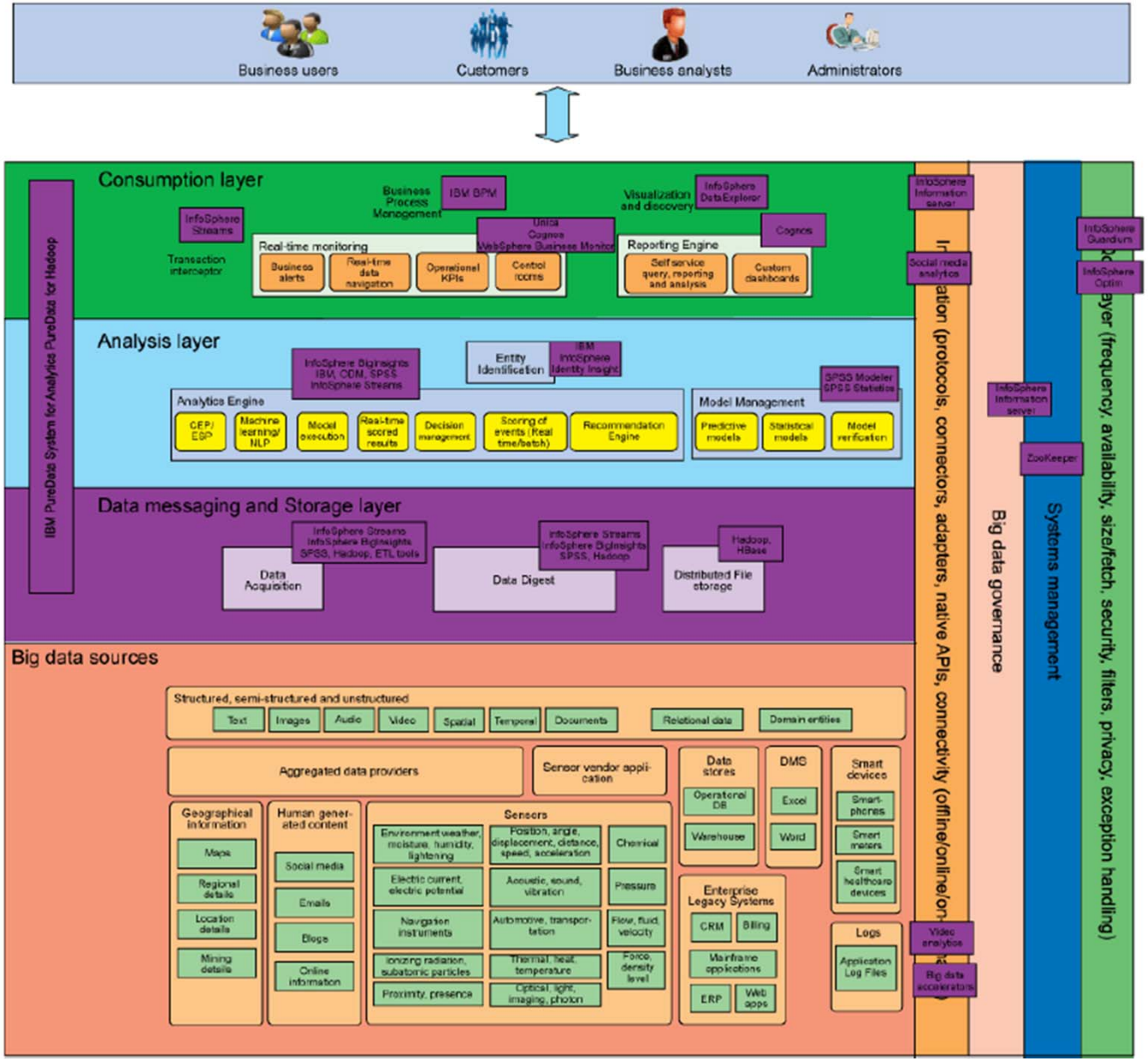
iv. Consumption Layer



iv. Consumption Layer

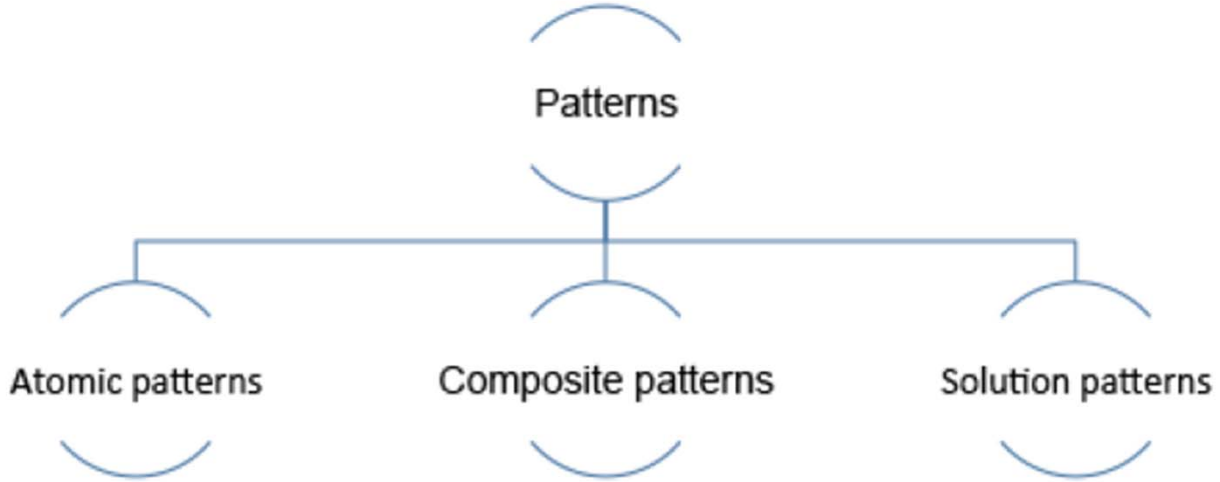
- consumes the output provided by the analysis layer
- consumers can be visualization applications, human beings, business processes, or services
- challenging to visualize the outcome of the analysis layer

Example: Insurance Fraud



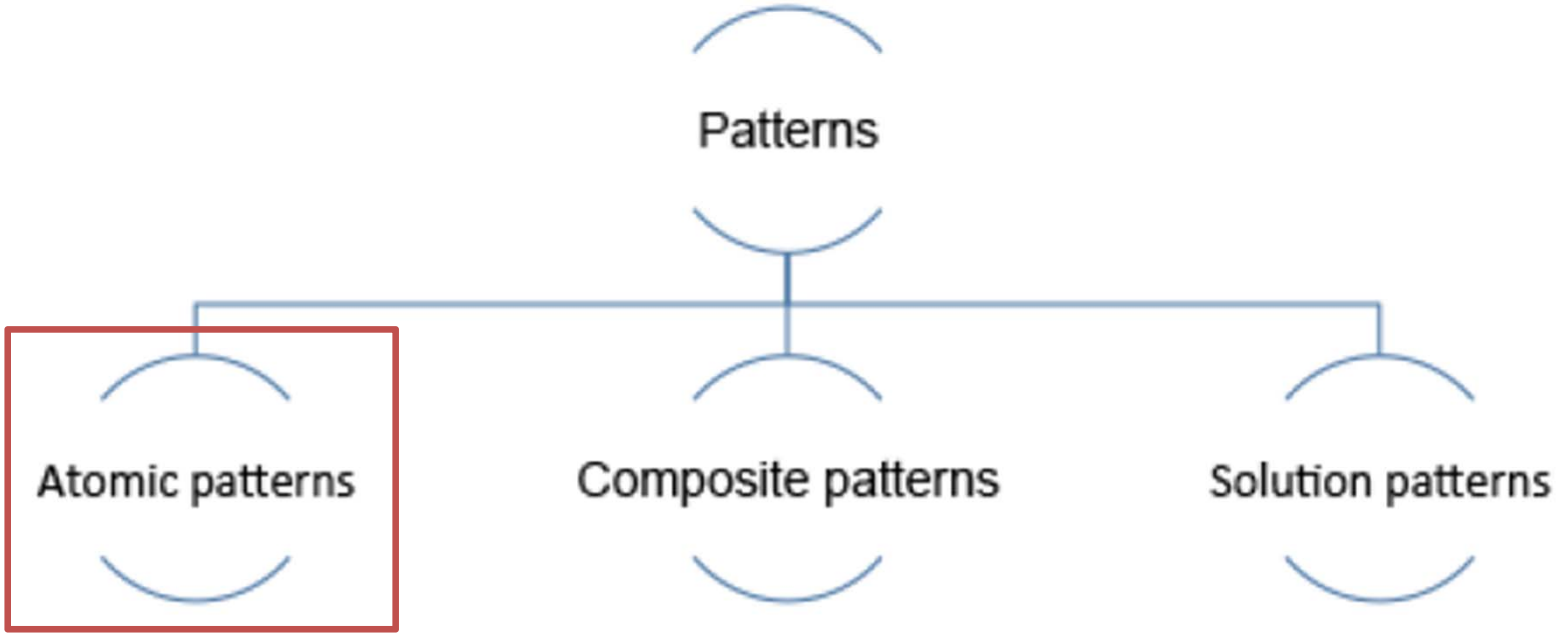
4. Understanding Atomic and Composite Patterns for Big Data Solutions

- Patterns help to define the parameter of a big data solution
- Patterns enable a structured approach to establish the scope and define the high-level solution for a big data problem
- Categories of patterns:

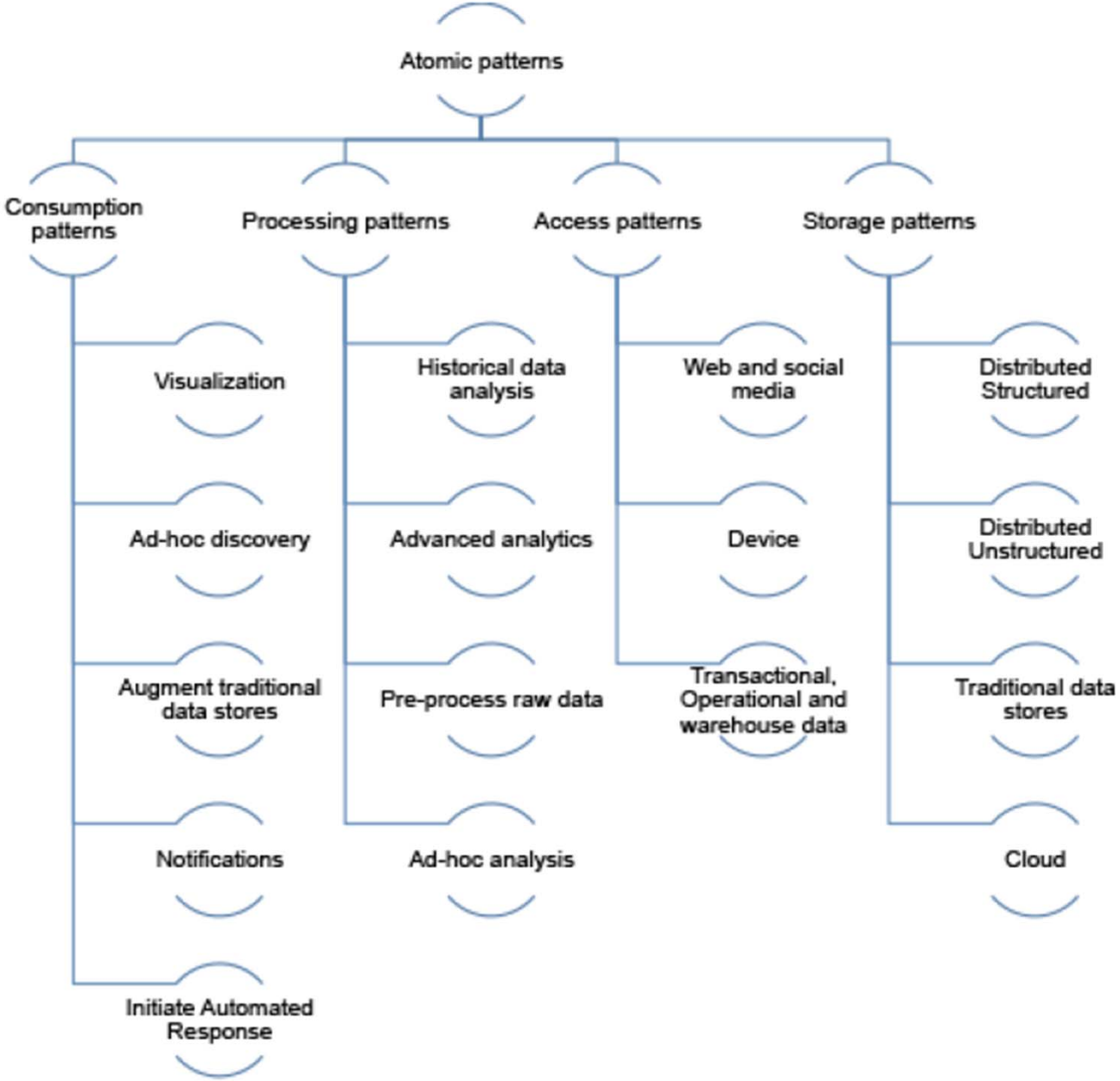


ATOMIC PATTERNS

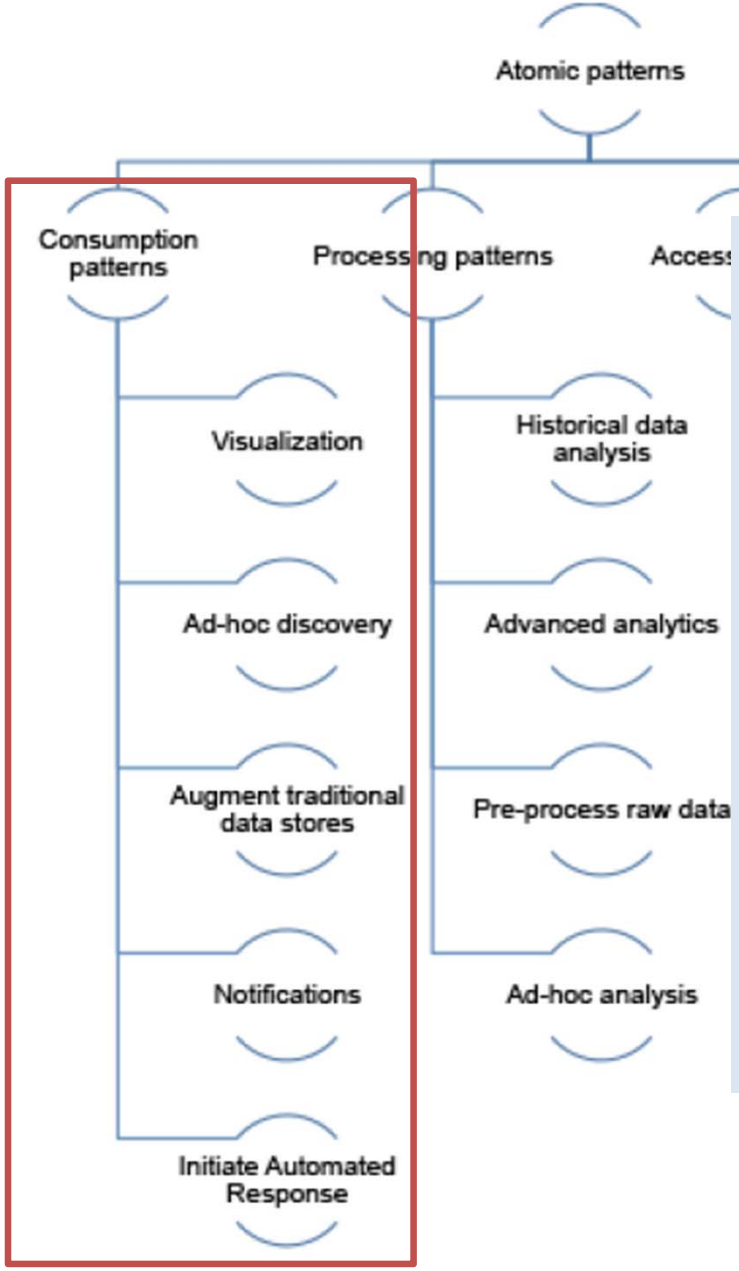
describe approaches for consuming, processing, accessing and storing big data



Types of Atomic Patterns

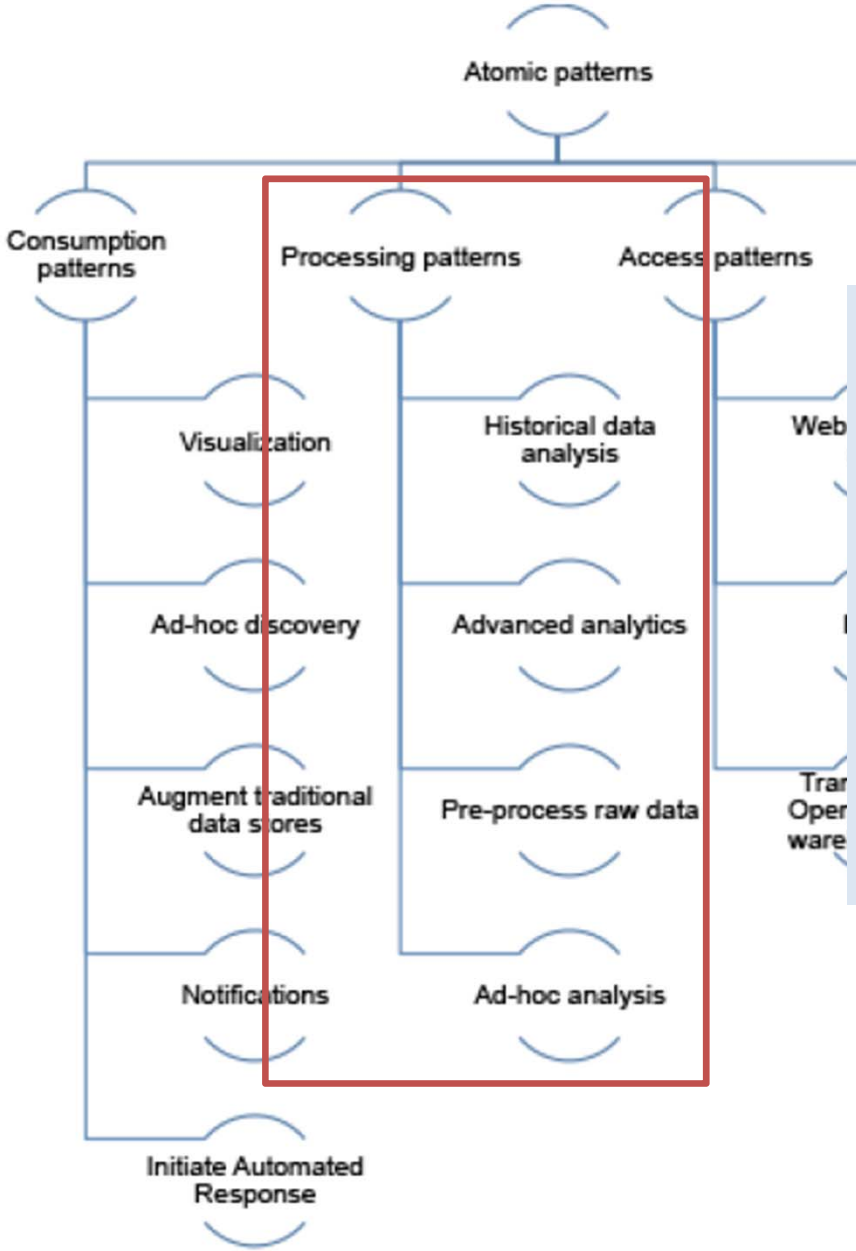


Atomic: Data Consumption Patterns



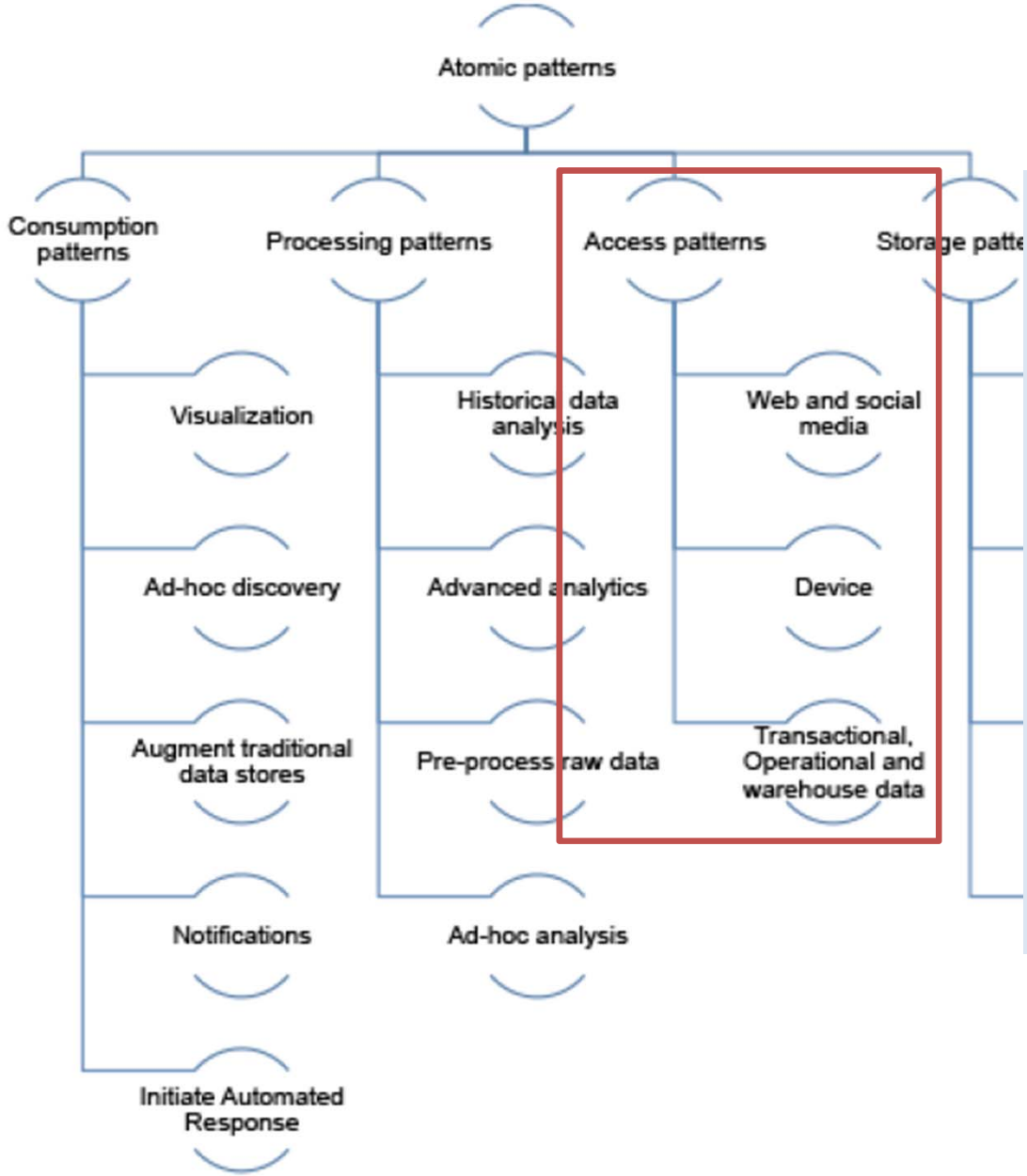
- Visualization Pattern
- Ad-hoc Discovery Pattern
- Augment Traditional Data Stores
- Notification Pattern
- Initiate an Automated Response Pattern

Atomic: Processing Patterns



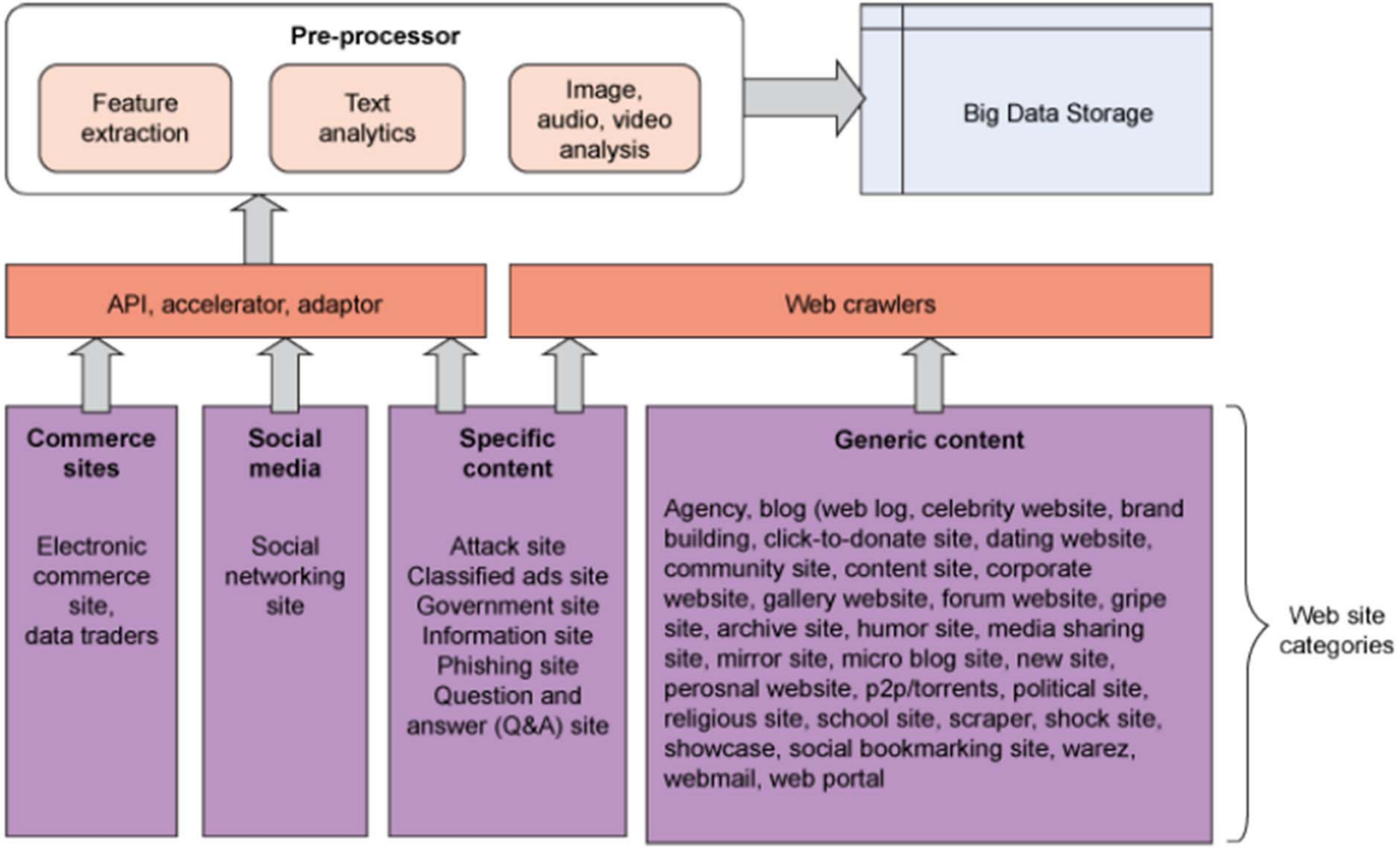
- Historical Data Analysis Pattern
- Advanced Analytics Pattern
- Pre-process Raw Data Pattern
- Ad-hoc Analysis Pattern

Atomic: Access Patterns

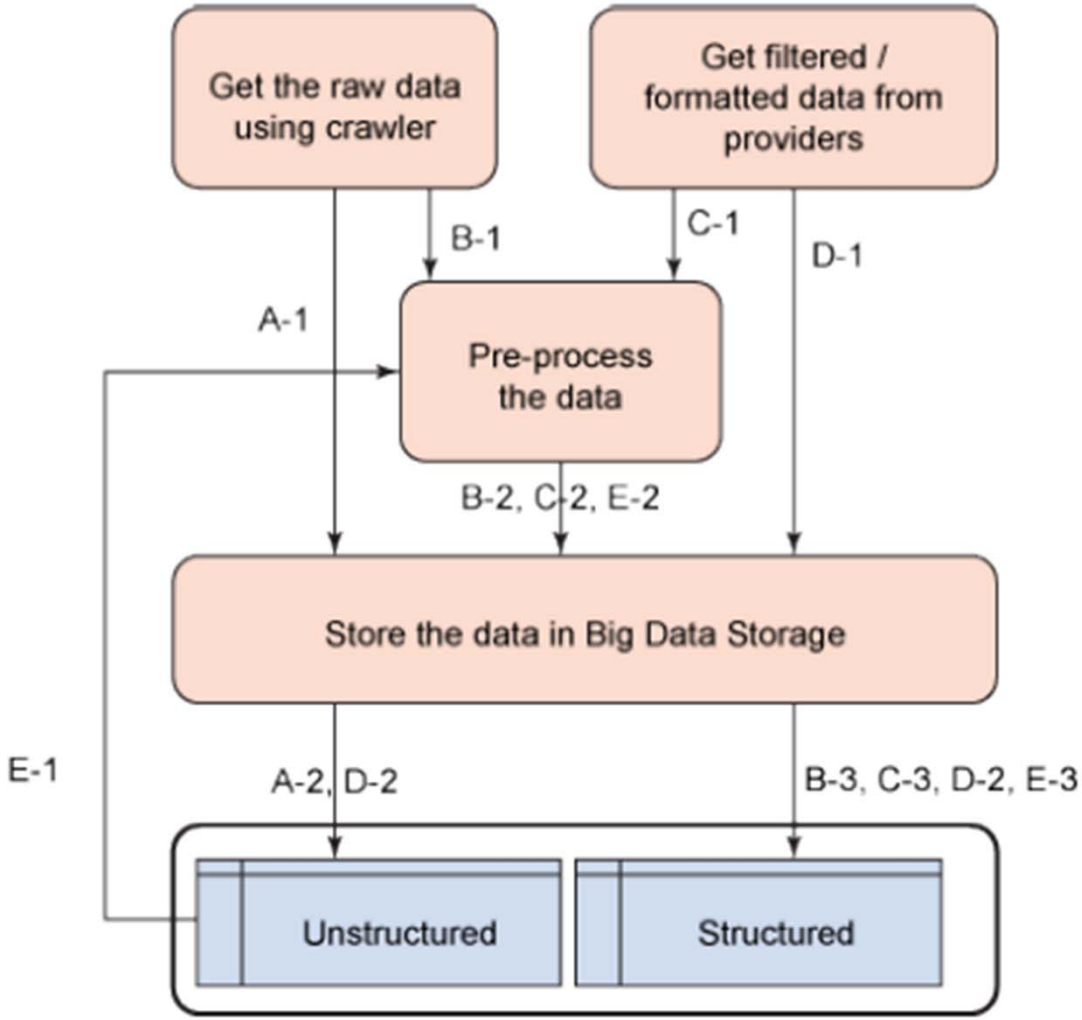


- Web and Social Media Access Pattern
- Device-generated Data Pattern
- Transactional, Operational and Warehouse Data Pattern

Access: Web and Social Media Access Pattern



Web and Social Media
Access Pattern: Big Data Accessing Steps



Big Data Accessing Steps

- **Web media access for data in unstructured storage**
  - Step A-1. A crawler reads the raw data
  - Step A-2. The data is stored in unstructured storage

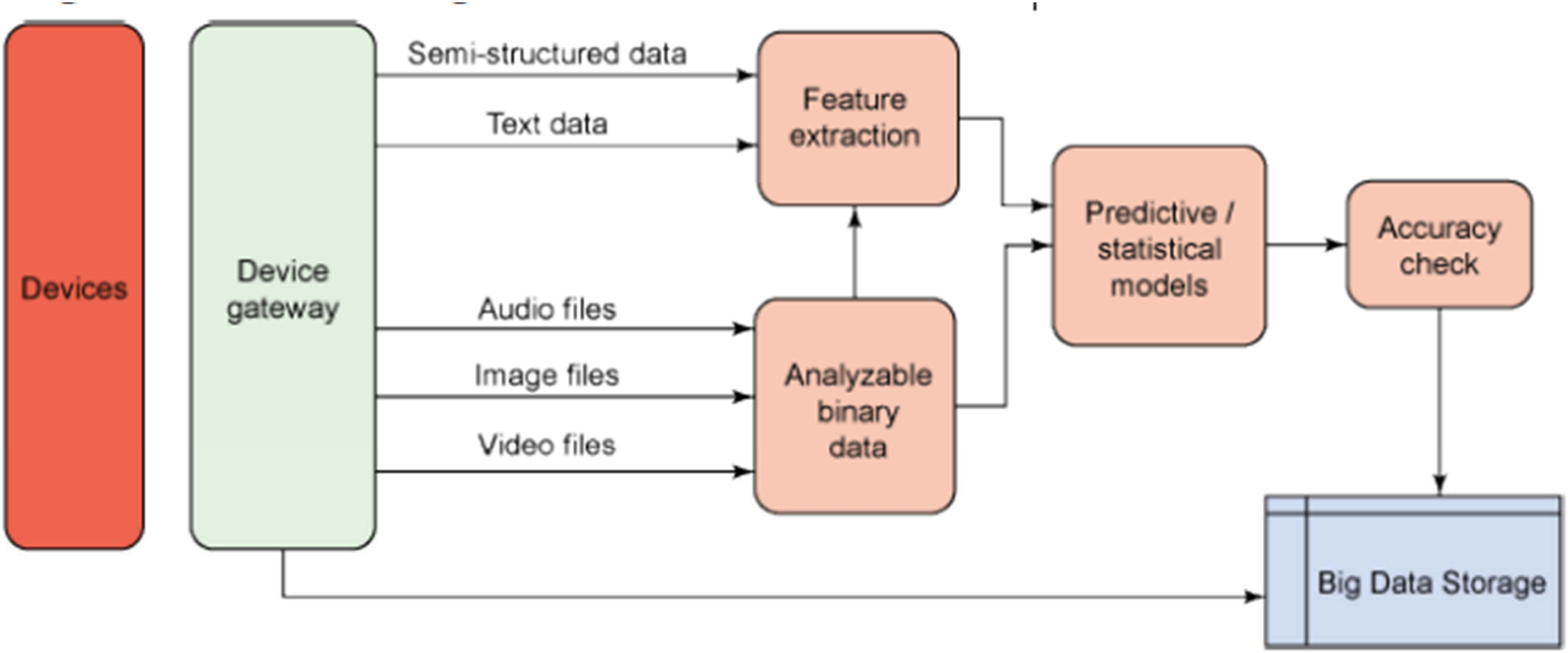
- **Web media access pre-process data for structured storage**
  - Step B-1. The crawler reads the raw data
  - Step B-2. This data gets pre-processed
  - Step B-3. The data is stored in structured storage

- **Web media access to pre-process unstructured data**
  - Step C-1. Data from the providers can be unstructured, in rare cases
  - Step C-2. Data is pre-processed
  - Step C-3. Data is stored in structured storage

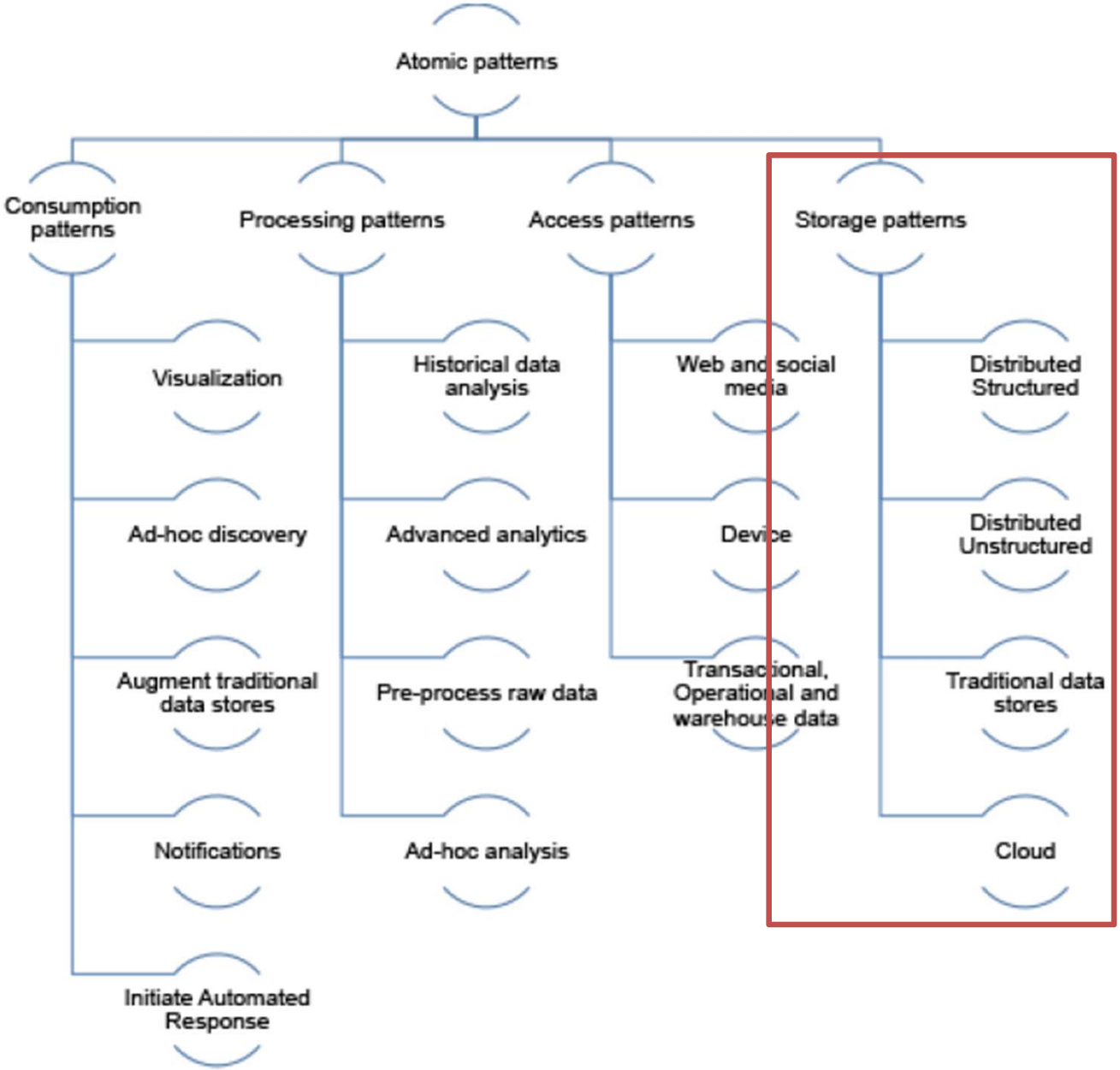
Big Data Accessing Steps

- **Web media access for unstructured or structured data**
  - Step D-1. Data providers provide structured or unstructured data
  - Step D-2. Data is stored in structured or unstructured storage
- **Web media access to pre-process unstructured data**
  - Step E-1. Unstructured data, stored without pre-processing, cannot be useful unless it is in a structured format
  - Step E-2. Data is pre-processed
  - Step E-3. Pre-processed, structured data is stored in structured storage

Device-Generated Data Access

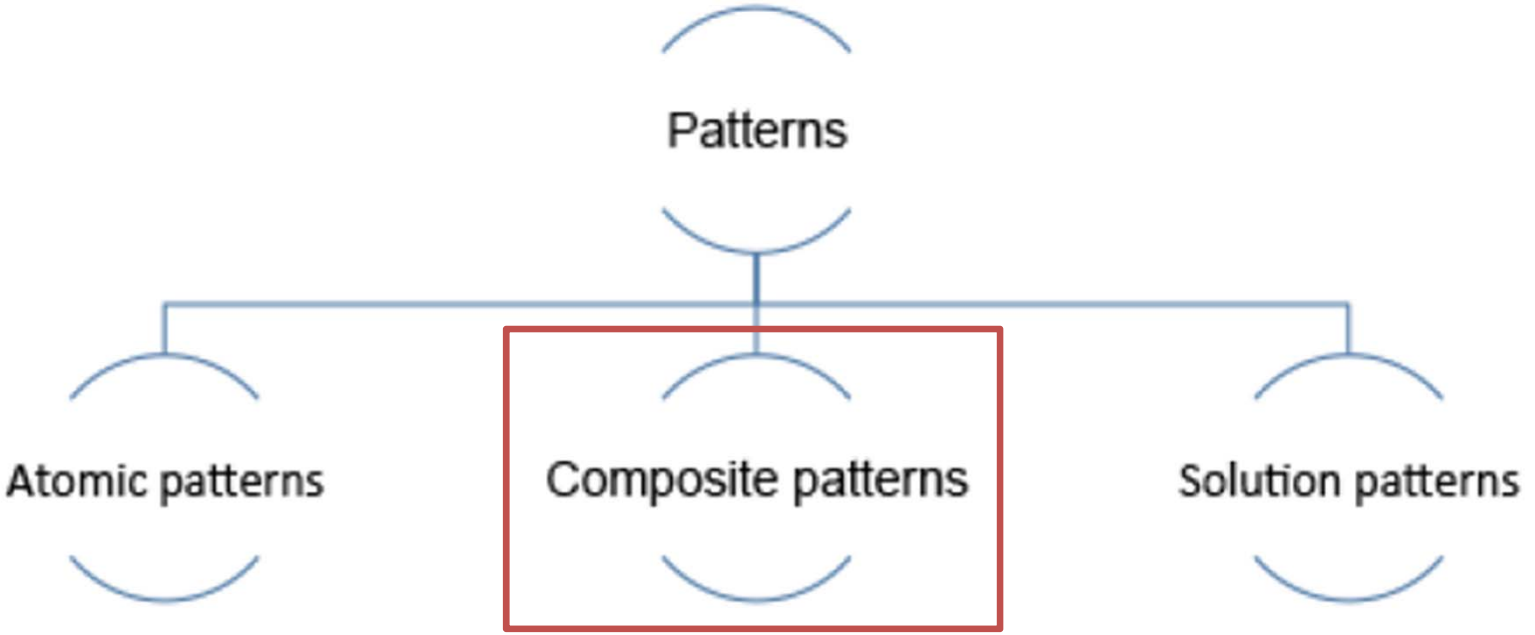


Atomic: Storage Patterns



- Distributed and Unstructured Data
- Distributed and Structured Data
- Traditional Data Stores
- Cloud Storage

COMPOSITE PATTERNS



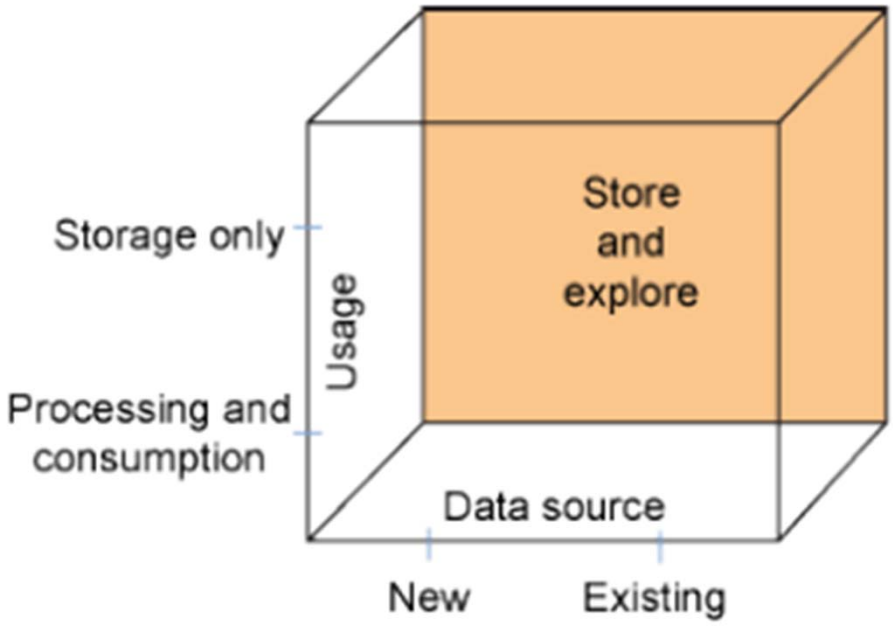
Composite Patterns

- *atomic patterns* focus on providing capabilities required to perform individual functions but composite patterns map to one or more atomic patterns to solve a given business problem

- classified based on the end-to-end solution and each composite pattern has one or more dimensions to consider

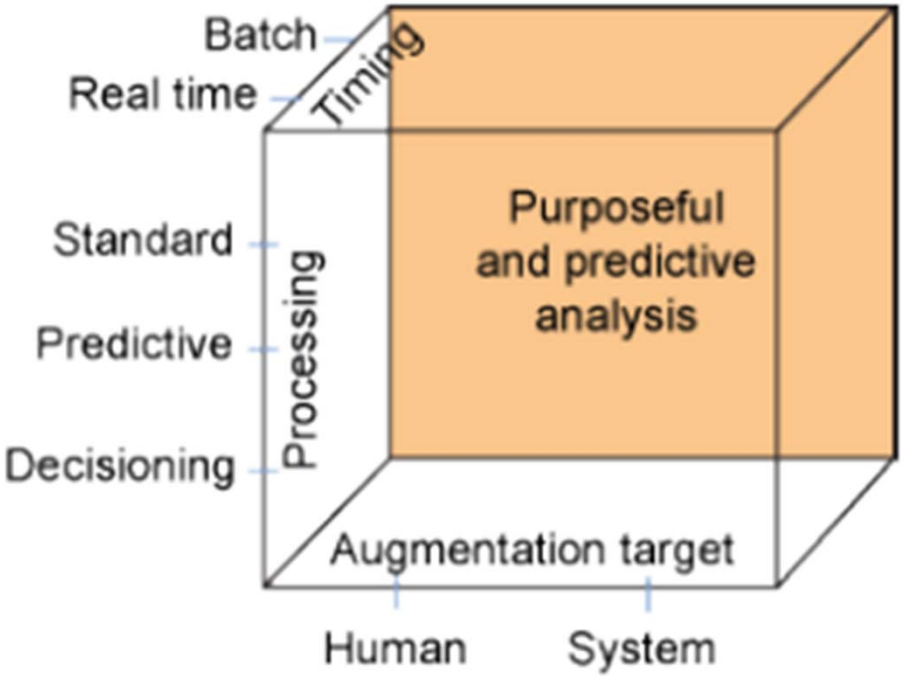
Store and Explore Composite Pattern

useful when business problem demands storing a huge amount of new and existing data that has been previously unused because of lack of adequate storage and analysis capability



Purposeful and Predictable Analysis Composite Pattern

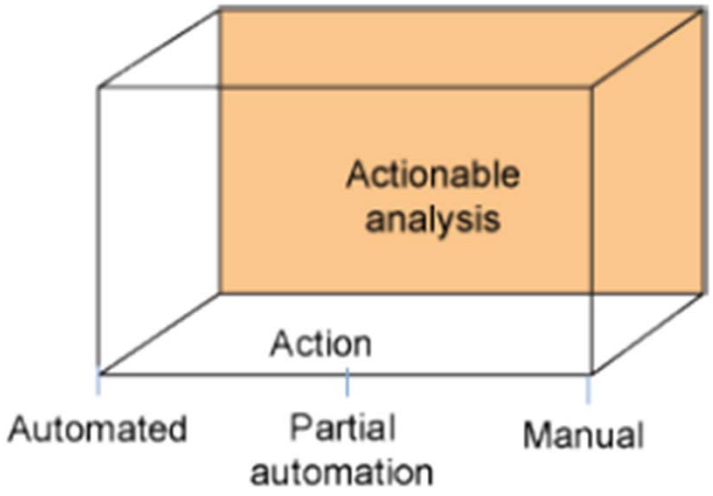
used to perform analysis using various processing techniques; may enrich existing data with new insight or create output that can be consumed by various users



Actionable Analysis Pattern

Analysis is performed on the set of data and actions are implied based on the repeatable past actions or on an action matrix

- Analyze the data to get the insight
- Make a decision
- Activate the appropriate channel to take the action to the right consumer



Composite to Atomic Patterns Mapping

Atomic Patterns →		Consumption patterns					Processing patterns				Access Patterns			Storage Patterns			
Composite Patterns ↓		Visualization	Ad-hoc discover	Augment traditional data stores	Notifications	Initiate Automated Response	Historical data analysis	Advanced analytics	Pre-process raw data	Ad hoc analysis	Web and Social media	Device	Transactional, Operational and warehouse data	Distributed-Unstructured	Distributed-Structured	Traditional data stores	Cloud
Store & explore																	
Data source	Existing												✓				
	New										✓	✓	✓				
Usage	Storage only													✓	✓	✓	✓
	Processing & consumption	✓	✓	✓			✓		✓	✓							
Purposeful and predictive analysis																	
Augmentation target	Human	✓	✓		✓												
	System			✓		✓										✓	
Processing	Standard						✓		✓	✓	✓	✓	✓	✓	✓		✓
	Predictive							✓	✓		✓	✓	✓	✓	✓		✓
	Decisioning							✓	✓		✓	✓	✓	✓	✓		✓
Timing	Batch			✓		✓	✓	✓	✓		✓	✓	✓	✓	✓		✓
	Real time	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓		✓
Actionable analysis																	
Action	Full automated			✓	✓	✓	✓	✓			✓	✓	✓	✓	✓		✓
	Partial automation	✓		✓	✓		✓	✓			✓	✓	✓	✓	✓		✓
	Full manual	✓		✓	✓		✓	✓			✓	✓	✓	✓	✓		✓

5. Apply a Solution Pattern to your Big Data Problem and Choose the Products to Implement it

1. **Getting started** – Store and Explore
2. **Gaining advanced business insight** – Purposeful and predictive analytics
3. **Take the next best action** – Actionable analysis

Example: Insurance Fraud

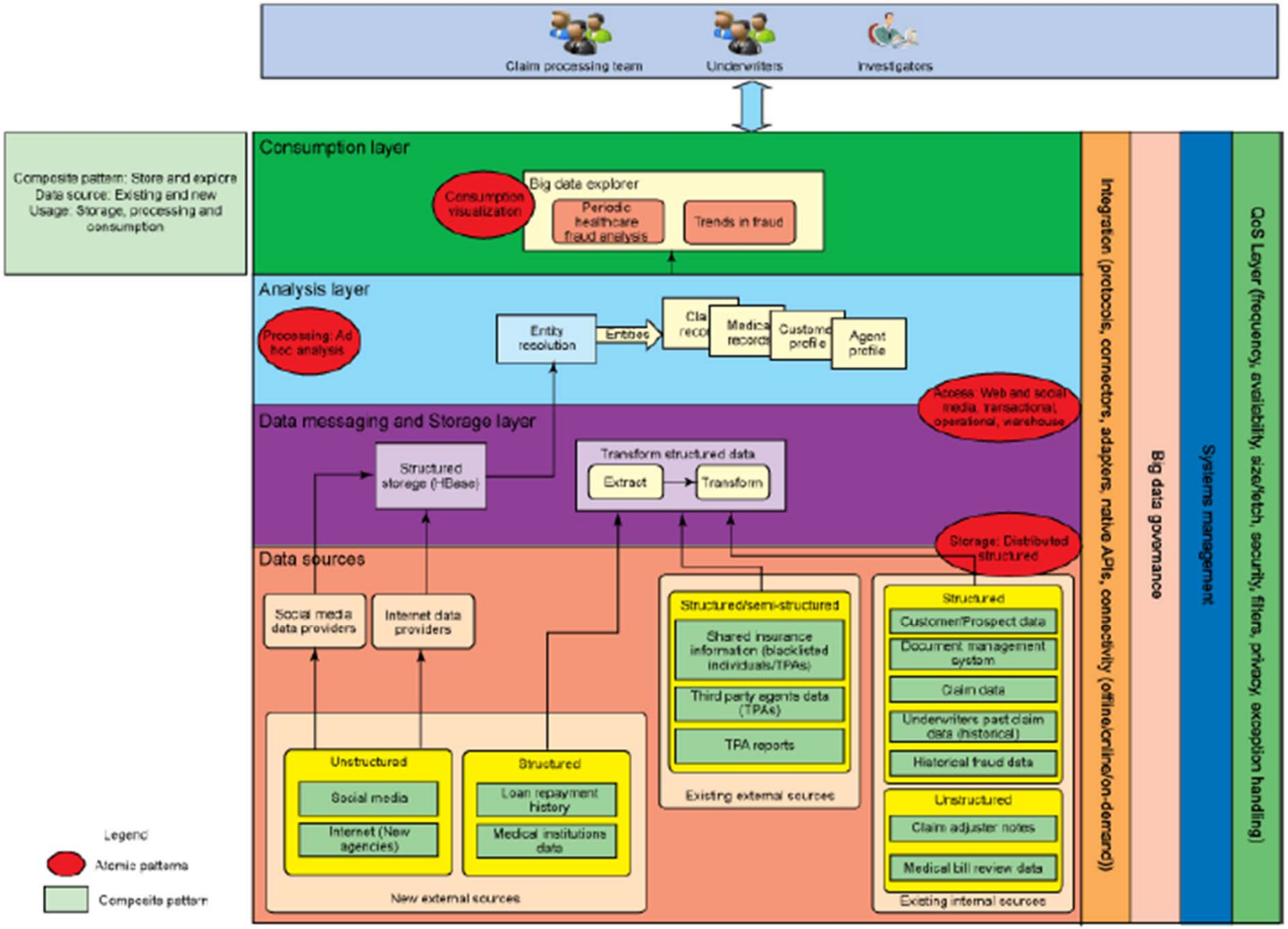
- Insurance fraud is an act or omission intended to gain dishonest or unlawful advantage, either for the party committing the fraud or for other related parties

- **Current fraud-detection process**
  - Monitoring fraud
  - Searching for potential fraud indicators
  - Coordinating with law enforcement agencies

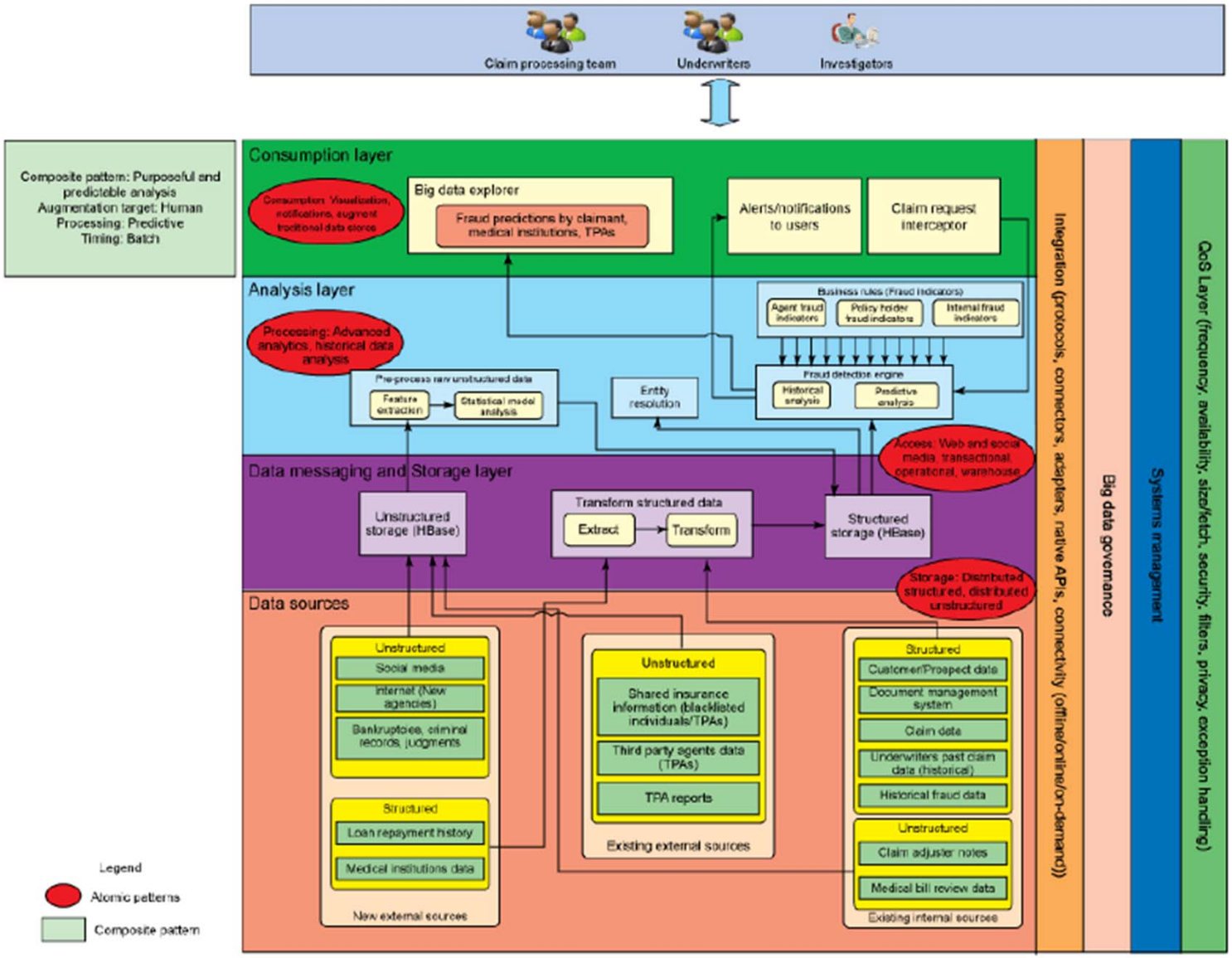
Example: Insurance Fraud

- **Issues with the current fraud-detection process**
  - Limited Data Sets
    - Traditional solutions use models based on historical fraud data, black-listed customers and insurance agents, and regional data about fraud peculiar to a certain area
  - Manual Work
    - Insurers may not be able to investigate all the indicators. Fraud is often detected very late, and it is difficult for the insurer to do adequate follow up for each fraud case
  - Hard to deal with new cases
    - Current fraud detection relies on what is known about existing fraud cases, so every time a new type of fraud occurs, insurance companies have to bear the consequences for the first time

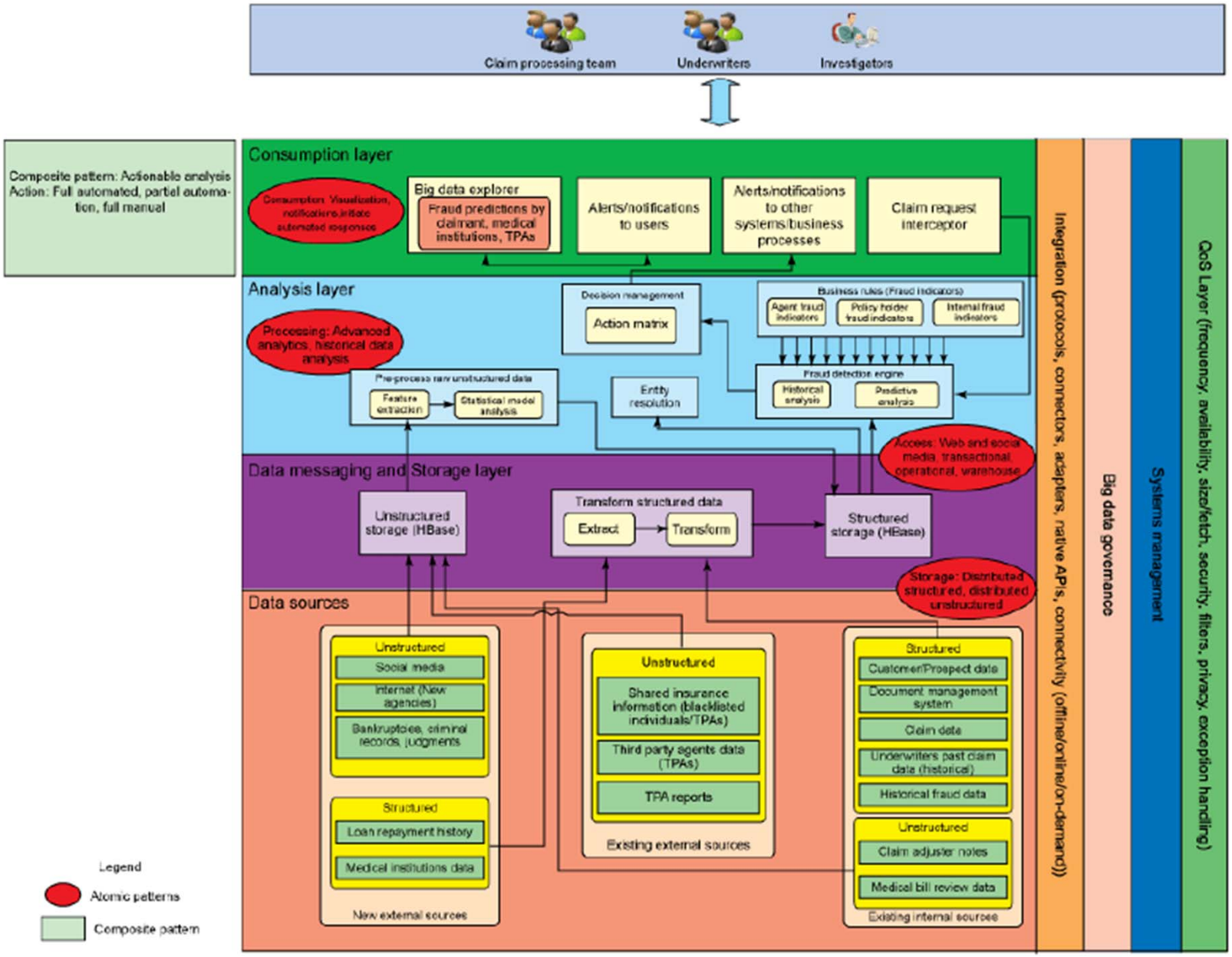
1. Solution Pattern: Getting Started



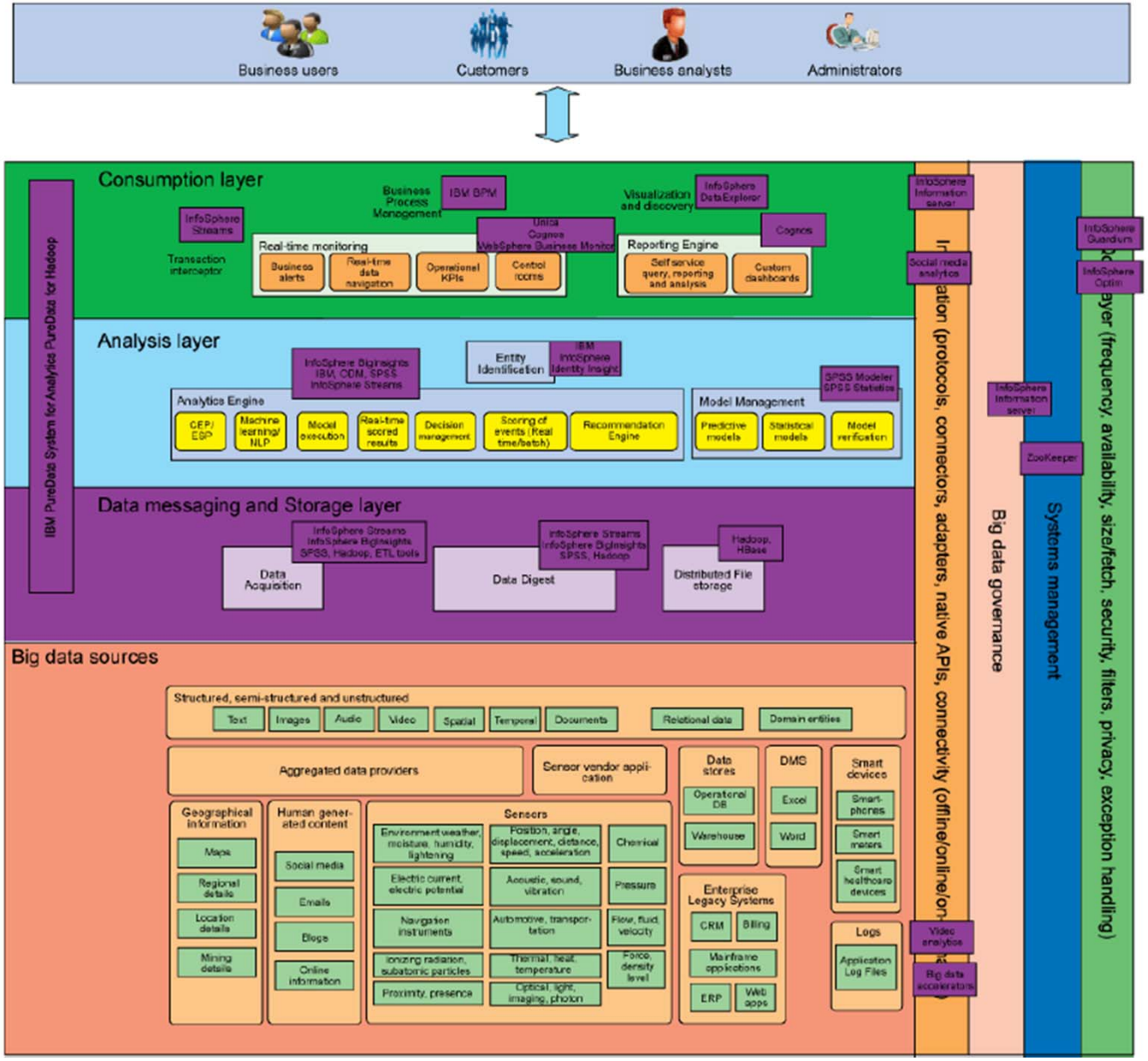
2. Solution Pattern: Gaining Advanced Business Insight



3. Solution Pattern: Take the Next-best Action



Products and Technologies Mapped to Logical Layers Diagram



Summary

1. Big data classification and architecture
2. What and how in a big data solution
3. 4 logical architectural layers in a big data solution
4. Patterns: atomic, composite and solution
5. Putting all together: problem to solution