Innovazione nell'analisi dei dati
8 - 15 maggio

Coordinatore: F. Stanco
DMI - Università degli Studi di Catania
M. Marroccia, G. Ursino, G. Scuderi, F. Milotta
STMicroelectronics
STMicroelectronics: a global presence

- One of the world’s largest semiconductor companies
- 2019 revenues of $9.56B
- 46,000 employees of which 7,800 in R&D
- Over 80 Sales & marketing offices serving over 100,000 customers across the globe
- 11 Manufacturing sites
- Signatory of the United Nations Global Compact (UNGC), Member of the Responsible Business Alliance (RBA)

Top 10 Customers 2019

<table>
<thead>
<tr>
<th>Position</th>
<th>Company</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Apple</td>
</tr>
<tr>
<td>2</td>
<td>Huawei</td>
</tr>
<tr>
<td>3</td>
<td>Bosch</td>
</tr>
<tr>
<td>4</td>
<td>Mobileye</td>
</tr>
<tr>
<td>5</td>
<td>Ciena</td>
</tr>
<tr>
<td>6</td>
<td>Samsung</td>
</tr>
<tr>
<td>7</td>
<td>Continental</td>
</tr>
<tr>
<td>8</td>
<td>Seagate</td>
</tr>
<tr>
<td>9</td>
<td>HP</td>
</tr>
<tr>
<td>10</td>
<td>Tesla</td>
</tr>
</tbody>
</table>
Affordable, desirable electric vehicles
Increase safety for road users & driver comfort and convenience
Cleaner, greener Internal Combustion Engines

Making **driving** safer, greener and more connected

Decrease carbon emissions to reduce global warming impact
Increase use of renewable energy

Making **homes & cities** smarter, for better living, higher security, and to get more from available resources

Enabling the evolution of **industry** towards smarter, safer and more efficient factories and workplaces

Cloud connected and data-enabled services
Digital security for all data
5G accelerating the connection of objects to the IoT

Making everyday **things** smarter, connected and more aware of their surroundings
We are drivers of your innovation

Advanced R&D centers around the world for close collaboration with operations and customers

- **18,500** patents & 590 new filings in 2019
- **16%** of revenues invested in R&D
- **7,800** people working in R&D and product design
Our technology starts with our people

<table>
<thead>
<tr>
<th>Manufacturing</th>
<th>~ 67%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research &amp; Development</td>
<td>~ 17%</td>
</tr>
<tr>
<td>Marketing &amp; Sales, Divisional Functions, Administration &amp; General services</td>
<td>~ 16%</td>
</tr>
</tbody>
</table>

As of December 31, 2019

<table>
<thead>
<tr>
<th>Region</th>
<th>Number of People</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italy</td>
<td>~10,600</td>
</tr>
<tr>
<td>Mediterranean</td>
<td>~4,700</td>
</tr>
<tr>
<td>Asia</td>
<td>~18,400</td>
</tr>
<tr>
<td>France</td>
<td>~10,100</td>
</tr>
<tr>
<td>Others</td>
<td>~750</td>
</tr>
</tbody>
</table>
Sustainability

Product compliance (Conflict-mineral free, RoHS)

Eco-design devices (power-efficient & low-carbon)

Raw materials

Social & Environmental programs (RBA code of conduct, Water, Waste, Energy & Climate Change, Chemicals)

Manufacturing

Usage

End of life

Responsible applications (planet-friendly & human-welfare)

Eco-friendly devices (power-efficient & low-carbon)

Product compliance (Material Declaration & ECOPACK®)
Where will we be in 5 years?
VUCA

- Volatile
- Uncertain
- Complex
- Ambiguous

Did you know?

- Top 10 In-Demand Jobs in 2018
  - Did not exist in 2008

- 1984: 1000 @ devices
- 2020: 75,000,000,000 @ devices

- The « PRESENT » has never been SO SHORT

- 5,500,000,000 searches on
  - Google per day
  - 400 hours of video per minute to
    - YouTube
  - 692,000,000 Tweets per day
  - 54,000,000,000 What’s app per day

- 2,200,000,000 Facebook Users
- 1,100,000,000 Instagram Users
- 33 zettabytes of new data created in 2018

- DISRUPTION is new « Normal »
- BUT
- CONSTRUCTION is ESSENTIAL

We are preparing STUDENTS
  - For jobs that do not yet exist...
  - For technologies that have not been invented...

“Imagination is more important than knowledge.”
A. Einstein
Innovazione nell'analisi dei dati
Cloud Architectures

venerdì 8 maggio - 09.00-12.00

Coordinatore: F. Stanco
DMI - Università degli Studi di Catania

M. Marroccia, G. Ursino, G. Scuderi
STMicroelectronics
CLOUD ARCHITECTURES - FOCUS ON SECURITY AND DATA GOVERNANCE ASPECTS

STMicroelectronics
Cloud data Governance - How to Successfully Design and Implement a Data-Centric Security Architecture

Giuseppe URSINO
# What’s “Cloud Computing”

# What’s “XaaS”

# What is Data Governance

# Data Integration

# Data Classification
What is Cloud Computing?

A style of computing in which scalable and elastic IT-enabled capabilities are delivered as a service using internet technologies.

1. Service-Based
Consumer concerns are abstracted from provider concerns through service interfaces that are well-defined. The interfaces hide the implementation details and enable a completely automated response by the provider of the service to the consumer of the service. In addition, the service could be considered “ready-to-use” or “off the shelf” because the service is designed to serve the specific needs of a set of consumers, and the technologies are tailored to that need rather than the service being tailored to how the technology works. The articulation of the service feature is based on service levels and IT outcomes (availability, response time, performance versus price, and clear and predefined operational processes), rather than technology and its capabilities. In other words, what the service needs to do is more important than how the technologies are used to implement the solution.

2. Scalable and Elastic
The service can scale capacity up or down as the consumer demands at the speed of full automation (which may be seconds for some services and hours for others). Elasticity is a trait of shared pools of resources. Scalability is a feature of the underlying infrastructure and software platforms. Elasticity is associated with not only scale but also an economic model that enables scaling in both directions in an automated fashion. This means that services scale on-demand to add or remove resources as needed.

3. Shared
Services share a pool of resources to build economies of scale. IT resources are used with maximum efficiency. The underlying infrastructure, software or platforms are shared among the consumers of the service (usually unknown to the consumers). This enables unused resources to serve multiple needs for multiple consumers, all working at the same time.

4. Uses Internet Technologies
The service is delivered using Internet Identifiers, formats and protocols, such as URLs, HTTP, IP and representational state transfer web-oriented architecture.

5. Metered by Use
Services are tracked with usage metrics to enable multiple payment models. The service provider has a usage accounting model for measuring the use of the services, which could then be used to create different pricing plans and models. These may include pay-as-you-go plans, subscriptions, fixed plans and even free plans. The implied payment plans will be based on usage, not on the cost of the equipment. These plans are based on the amount of the service used by the consumers, which may be in terms of hours, data transfers or other use-based attributes delivered.
When it has made the decision to consider cloud services for an application or infrastructure deployment, it’s important to grasp the fundamental differences between the core categories of cloud services available.

The cloud is a very broad concept, and it covers just about every possible sort of online service, but when businesses refer to cloud procurement, there are usually three models of cloud service under consideration, Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). Each has its own intricacies and hybrid cloud models, but today we’re going to give you an understanding of the high-level differences between SaaS, PaaS, and IaaS.

**Software-as-a-Service (SaaS)**
Software deployment model where applications are delivered as a service to the customer
- **CRM as a service** (e.g., Salesforce)
- **Office Productivity Software** as a service (e.g., O365)

**Platform-as-a-Service (PaaS)**
Software development platform that supports the full software lifecycle (design, testing, and development)
- **Development & Testing** as a service (e.g., Microsoft Azure)
- **Database** as a service (e.g., Azure SQL DB, Azure Hyperscale SQL DB)

**Infrastructure-as-a-Service (IaaS)**
Model for provisioning core computing power (servers, storage) for applications and data, virtualization, etc. (provisioning of VM, containers, etc.)

‘XaaS’ is a style of computing where scalable and elastic IT-related capabilities are provided “as a service” to external customers using internet technologies.
WHAT is Data Governance?

**Set of roles**, accountabilities and responsibilities within the organization to define and orchestrate the corporate guidelines, processes, policies, standards, and technologies implementations.

**Policies** definition and orchestration

**Usage** of data to foster Company’s Strategy

Management, ownership, security, availability, accessibility, quality, consistency, and auditability of data and their alignment with company strategy.
Protecting data has long been the primary goal for security organizations, as evidenced by the classic confidentiality, integrity and availability (CIA) triad. Yet, security controls focus on users and systems, and are rarely architected to focus on the data itself.

Data is an asset class that is difficult to secure comprehensively, unless a broader view of its life cycle is taken. The challenge organizations are facing today is that, unlike a server, data is pervasive and does not stay in a single silo. Data has become an infinitely mutable and movable resource that is stored and processed in a decentralized ecosystem of data silos, often for many different business purposes.

This semi-chaotic environment can frequently lead to the implementation of duplicate security controls in different silos or, worse, within a single environment. Besides the financial and operational implications, this situation can introduce inconsistencies in control outcomes, leading to a lower level of risk management.
HOW? Set of role in Data Governance

• **Data Owner (Business representative)**
  - Executive manager who has responsibility over the business domain
  - Responsible of the data owned by his/her business organization, he/she leads and promote data governance to ensure data security, availability and quality
  - Appoint the data steward(s)

• **Data Steward (Business representative)**
  - Expert of Business Domain data
  - Ensure the right data usage, computation, data process, monitoring, data generation through business processes by providing correct data requirements
  - Define the policies and the guidelines to preserve data value, quality, security, accessibility and ensure data governance aligned with corporate strategy
  - Can delegate one or more Data Stewardships for dedicated area/programs

• **Data Stewardship**
  - Same as Data Steward for particular initiatives and/or area.

• **Data Gatekeeper/Custodian**
  - DIT Expert of Business Domain
  - DIT counterpart of the Data Steward
  - Support Data Steward by ensuring implementation of all the requirements and policies to properly manage and integrate data in the IT Landscape
**Data Owners** are business people, usually executive profiles, who have direct line responsibility for the functional area that owns the data.

- **Accountabilities:**
  - Implementation of data governance organization
  - Definition and implementation of processes, guidelines and standards
  - Data compliance to external / internal regulation (GDPR, liability, SOP, contractual liabilities…)
- **Responsibilities:**
  - Nominate Data Steward
  - Approval of exceptions

As leaders in the user community, they are part of the team that drives the governance process since the need for governance should originate and be maintained in the business community.

**Data Steward** is a business senior manager, domain’s data expert, that can influence business and operational decisions, and obtain stakeholder commitments.

- **Accountabilities:**
  - Ensuring monitoring of information assets and data collection
  - Data Security, Protection, Quality, Consistency, and Availability
  - Ensuring data access controls according to corporate policies and compliance to regulation
  - Monitoring usage, relevance and data quality (accuracy, completeness, consistency, timeliness,…..) of data published
- **Responsibilities:**
  - Ensuring data access controls according to corporate policies and compliance to regulation
  - Definition and implementation of processes, guidelines and standards
  - Data Compliancy to external / internal regulation (GDPR, liability, SOP, contractual liabilities…)
  - Data Security, Protection, Quality, Consistency, and Availability
  - Ensuring monitoring of information assets and data collection

**Data stewards rely on a support network within the business organization to deploy processes, ensure data accuracy, timeliness, analysis and resolution of issues or to coordinate the same with external partners, customers and vendors when required.**

**Data Gatekeeper/Custodian** ensures the technology support to data governance; collaborating with Data Steward and implement systems, integration, and monitoring for data management.

- **Accountabilities:**
  - Data analysis and provide proper artifacts to Data Steward
  - Implementation of given policies, guidelines, and standards
- **Responsibilities:**
  - Support Data Steward by implementing monitoring usage, relevance and data quality (accuracy, completeness, consistency, timeliness,…..) of data published
  - Granting data governance as per given policies implementing
    - Data Collection and maintenance
    - Data integration and processes
    - Data security
    - Access control
    - Data quality controls
  - Create and manage the metadata for published data sources ensuring usability and scalability
  - Trigger Data Governance Office and Data Steward when required
Data Governance Process

BUSINESS request
- BDM
- Ad Hoc (*)

Data Architecture Validation

Data Analysis

Data Governance Office
- Data Steward
- RMIS expert
- Data gatekeeper
- Enterprise Architect

Identification of Data Rules & Data Catalog scope

Risk Index evaluation + associated mitigation plan

Operational Implementation

Data Governance Office
- Data Analyst(s)
- Data Scientist(s)
- Business Analyst(s)

- Data Owner
- Exceptions
- Chief Data Officer

- Legal Data Steward

Find alternative solution

Choice approved

Implementation of solution

N

Y
Why Data Governance?

- Data is always changing
- Data grows exponentially
- Business is always changing
- Regulatory, Compliancy, Policy etc…
- New Security Policies
- New Risks

Common questions on and with the Cloud:
Are you still doing the right things?
Which Architecture(s)? Which Governance?
Key challenges for data governance and security

- Continual increases in the **Volume, Velocity, Variety** and business value of data.
- The increasing tension between the various regulatory requirements for data security and privacy and the business requirements for agility, flexibility, sharing, deep insight and partnering with other organizations.
- The need to automate and **scale data processing using artificial intelligence (AI) and machine learning (ML)** techniques.
- The decentralization of storage into an ecosystem of distinct and unrelated data silos, where traditional system-centric approaches to data security lose their efficacy.
- The increasing number and complexity of privacy regulations and laws requiring mandatory compliance, based on the Organization for Economic Co-operation and Development’s eight privacy principles. Examples include the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), CLOUD ACT.
- An upward trend in the number and scale of data breaches, throughout 2018 and 2019, that may have been subject to regulatory response, such as punitive fines.
2019 Cost of a Data Breach Report

Data breaches can cause devastating financial losses and affect an organization’s reputation for years.

From lost business to regulatory fines and remediation costs, data breaches have far reaching consequences. The annual Cost of a Data Breach Report, conducted by the Ponemon Institute and sponsored by IBM Security, analyzes data breach costs reported by 507 organizations across 16 geographies and 17 industries. Read the report to discover all the factors that influence the cost of a data breach and which security measures can help organizations reduce the financial impact.

**USD 3.92 million**
Average total cost of a data breach

**Sensitive Data is a Risk**

70%
Of organizations surveyed use live customer data in non-production environments (Development, testing, Q/A)

*Database Trends and Applications. Ensuring Protection for Sensitive Test Data*

**25,575 records**
Average size of a data breach

**United States**
Most expensive country: **USD 8.19 million**

52% of surveyed organizations outsource development

50% Of organizations surveyed have no way of knowing if data used in test was compromised
# The Four Data-Centric Control Families, With Examples of Commonly Implemented Controls

<table>
<thead>
<tr>
<th>DCSA Control Families</th>
<th>Control Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insight</td>
<td>Data Mapping</td>
</tr>
<tr>
<td></td>
<td>Data Discovery</td>
</tr>
<tr>
<td></td>
<td>Data Classification</td>
</tr>
<tr>
<td>Confidentiality</td>
<td>Access Control</td>
</tr>
<tr>
<td></td>
<td>Data Masking and Encryption</td>
</tr>
<tr>
<td></td>
<td>Enterprise Digital Rights Management (EDRM)</td>
</tr>
<tr>
<td></td>
<td>Data Loss Prevention (DLP)</td>
</tr>
<tr>
<td>Monitoring and Response</td>
<td>Security Information and Event Management</td>
</tr>
<tr>
<td></td>
<td>Database Activity Monitoring</td>
</tr>
<tr>
<td>Third-Party Governance</td>
<td>Contractual Controls</td>
</tr>
</tbody>
</table>
Coupling Data (Security) Governance With the Data Centric Security Architecture
How to Successfully Design and Implement a Data-Centric Security Architecture (DCSA)

**Prework: Understand the Purpose, the Elements and the Terminology**
- Data-centric control families
- Infrastructure-level controls
- Data- or object-level controls
- Data silos
- DCSA vs. data security governance (DSG)
- Coupling the DSG and DCSA

**Step 1: Data Mapping**
- Business expectations:
  - Data categories
  - Data locations
  - Business process
- Questions to ask:
  - What?
  - Where?
  - Why?
  - How?

**Step 2: Data Discovery and Classification**
- Find business realities:
  - Define classification scheme
  - Create classification metadata
- Find unprotected sensitive data
- Trigger protection measures

**Step 3: Data Flow Modeling**
- Create data flow model for key enterprise use cases
- Gain understanding of what is being modeled
- Avoid undue complexity
- Review and update periodically

**Step 4: Data Control Examination**
- What controls are available?
- Where are they applied?
- How does the data need to be protected?

**Step 5: Product Examination**
- Questions:
  - What data-centric security products are in use?
  - Which control families do they provide?
  - Which silos do they cover?
- Actions:
  - Identify actual and unused product capabilities
  - Determine maturity of product use
  - Rationalize product mix
  - Explain controls in the DFSG
  - Document open questions

---

**Data discovery and classification.** Find the data in more detail, record what is found, and codify its level of sensitivity and criticality to the organization. Assign owners to the data.

Describe how and why data moves through the system.

Examine data flows for typical data use cases. What DCSA controls are applied to data in each silo, and are they sufficient for both present operations and any known future plans?

Examine the security products you have in place:
- Which of the four DCSA control families do they address, and for what silo?
- What products can easily be used across more silos?
- What critical gaps must be alleviated by acquiring new technologies and products?
# Data Security, Data Compliance, Data Risk

## Data Security
- Access control
- Encryption
- Data Integrity
- Data Leakage Prevention

## Data Compliance
- Data Access Policy
- Data Retention
- Forensics

## Data Risk
- Preparing for Potential issues
- Data loss
- Data Inaccessibility (service outage)
- Data exposure

---

Do not forget!!!
Zones, Actors, Silos and Dataflows for Unstructured Data

Data Sources Outside the Enterprise
- IoT Sensor
- Data Subject
- Any Device

Incoming Data Flow

Enterprise-Managed Endpoints and Software
- Actor: Enterprise Employee on Enterprise-Owned Device
- Actor: Interacting Application
- Silos: Database, Unstructured Data Store

Internal Data Flow

Data Silos (Partially) Beyond the Control of the Enterprise
- Cloud-Based Object Storage
- Any Device

Outgoing Data Flow
## Product Capability Matrix

<table>
<thead>
<tr>
<th>Technical DCSA Control Families</th>
<th>Data Stores (Silos)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insight</td>
<td>Confidentiality</td>
</tr>
<tr>
<td></td>
<td>Activity Monitoring</td>
</tr>
<tr>
<td></td>
<td>Structured data</td>
</tr>
<tr>
<td></td>
<td>Unstructured data</td>
</tr>
<tr>
<td></td>
<td>Hadoop (H)</td>
</tr>
<tr>
<td></td>
<td>Unstructured data in SaaS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Products</th>
<th>Structured data (relational databases)</th>
<th>Hadoop (H)</th>
<th>Big data in SaaS, such as O365, Salesforce or ServiceNow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product 1</td>
<td></td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>Product 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Product 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Integration Architecture
8 assessment criteria to guarantee data integration

- **Ability to adapt** — How much of a solution can still be used if (or when) one or more of the applications being integrated need to be replaced, such as when you replace an on-premises application with a SaaS application?
- **Ability to extend** — How easy it is to make changes in the integration when required, such as adding fields to an integration between an ERP and an e-commerce application?
- **Ability to monitor** — How much visibility does this solution provide into integration uptime, performance and quality, such as integration with existing monitoring and alerting systems or built-in monitoring capabilities?
- **Ability to reuse** — How easy it is to reuse this solution for additional integrations, including reuse of artifacts such as business logic or data mapping? For example, can you reuse a product data integration between an ERP and an e-commerce system to share that data with search engines?
- **Security and data governance** — How much work is required to secure the solution, and how much control does it afford over sensitive or proprietary data? For example, does the integration functionality run on hardware you control?
- **Support for ad hoc integrators** — How easy it is for staff members who are not integration specialists to implement or update integrations? For example, some developers may need to create integrations in order to deliver a project.
- **Support for mediation** — How well does the solution support connecting endpoints with disparate styles, such as connecting an API to a messaging system, or converting XML to JSON?
- **Support for reliable communications** — How well does the solution deal with intermittent or unreliable connections between endpoints, such as with equipment in the field or mobile devices?
Some Possible Reasons for Integrations to Break

- Firewall rules changed
- Expired certificate
- Disk full
- Change to API
Point-to-Point Integrations Scale Poorly

Point-to-Point Integrations Create \((N \times M)\) Connections

Mediated Integrations Create \((N + M)\) Connections
Mediation Enables Integration and Guarantee proper Governance and Security
Example of Hybrid Integration Platform
Data Classification
Data classification is "the process of organizing information assets with an agreed-on categorization glossary, enabling effective and efficient prioritization for information governance policy spanning quality, security, access, privacy, storage and retention."
The Importance of data visibility and Classification

Organizations increasingly understand the need to better manage their data in order to extract value and justify the costs of security, collection and storage.

Regulations such as the European General Data Protection Regulation (GDPR) are requiring:

- Robust and risk-based data governance and security controls.
- Driving a need for better visibility of data
  Some data may be so sensitive so to be considered - CLASSIFIED as secret,

Being able to find, analyze and classify data is increasingly important to drive business value, support governance and achieve compliance.

The figure illustrates the business functions that benefit from good data insight, including visibility and classification.
### Common Drivers for Data Governance and Data Classification

<table>
<thead>
<tr>
<th>Source</th>
<th>Example</th>
<th>Applicable Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Privacy regulation</td>
<td>European GDPR</td>
<td>Processing of Personal data and rules relating to the free movement of personal data subjects in the EU</td>
</tr>
<tr>
<td>Cybersecurity regulation</td>
<td>New York State Department of Financial Services 23 NYCCR 500:</td>
<td>Consumers’ private data</td>
</tr>
<tr>
<td></td>
<td>Cybersecurity Requirements for Financial Services Companies</td>
<td></td>
</tr>
<tr>
<td>Health and insurance</td>
<td>Health Insurance Portability and Accountability Act (HIPAA)</td>
<td>Protected health information (PHI)</td>
</tr>
<tr>
<td>Contract</td>
<td>IT or business process outsourcing. Cloud Service Provision (SaaS or</td>
<td>Any data shared or accessible as a result of the contract agreement</td>
</tr>
<tr>
<td></td>
<td>infrastructure as a service [IaaS])</td>
<td></td>
</tr>
<tr>
<td>Financially sensitive</td>
<td>Merger and acquisition data</td>
<td>Target acquisition information, including the fact of the existence of the intent</td>
</tr>
<tr>
<td>information</td>
<td>Corporate performance</td>
<td>Financial performance report prior to release to stock market</td>
</tr>
<tr>
<td>Intellectual property</td>
<td>Engineering or research records</td>
<td>Any data deemed by the organization to contribute to the competitive advantage, revenue and profitability of the organization</td>
</tr>
<tr>
<td></td>
<td>Process documentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Exploration data</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Customer data and sales opportunities</td>
<td></td>
</tr>
<tr>
<td>Security data</td>
<td>Security logs</td>
<td>Information material to the protection of the organization, including the organization’s technical infrastructure</td>
</tr>
<tr>
<td></td>
<td>Enterprise architecture (EA)</td>
<td></td>
</tr>
</tbody>
</table>
Data classification in the data life cycle

- Data Inception
  - Data Creation
  - Data Collection
- Legacy Data

- Discover and Classify

- Data Processing and Use
  - Classify
  - Control
  - Report

- Data Deletion

Visibility
Data classification tools and their relation to the data life cycle

CASB = Cloud Access Security Broker; DAG = Data Access Governance; DCAP = Data Centric Audit and Protection; EDRM = Enterprise Digital Rights Management; FA = File Analysis; DLP = Integrated Data Loss Prevention; UDC = User Driven Classification
O365 Data Security Native Tools

**Basic Security**
- Anti-Malware
- Anti-Spam
- SharePoint Anti-Malware
- DMARC/DKIM/SPF

**Advanced Security**
- Threat Protection (Threat Explorer, Tracker, Automated Investigations, Attack Simulator)
- O365 Cloud App Security/Microsoft Cloud App Security
- Advanced Threat Protection (Safe Links, Safe Attachments, Anti-Phishing)
- Message Encryption or S/MIME
- Azure RMS (MS Key/BYOK/New HYOK vs Old AD RMS (HYOK))
- Customer Key (Service Level)

**Encryption**
- Azure Information Protection: Data Classification
- Data Loss Prevention
- Advanced Data Governance

**Data Protection**
- Azure Information Protection: Data Classification
- Data Loss Prevention
- Advanced Data Governance

**Compliance**
- Compliance Manager (Compliance Score)
- Communication Compliance (Supervision), Insider Risk Management

**Management**
- Microsoft Secure Score
- O365 Management Activity and Service Health API

**Identity Access Management**
- Azure AD Identity Protection
- Azure AD Privileged Identity Management/Privileged Access Management
- Azure Advanced Threat Protection
- Multifactor Authentication
- Conditional Access (Intune, MCAS, Azure AD PIM)

**Audit and Logging**
- Advanced E-Discovery
- Customer Lockbox
- Audit Log Search

**Management Consoles**
- Exchange
- Skype for Business
- SharePoint
- OneDrive
- Yammer
- Power Apps
- Flow
- Security and Compliance
- Azure AD
- Intune

**Azure Sentinel**
- MCAS Console
- Azure Sentinel Console

**Microsoft Endpoint Manager (Intune + System Center)**
- EMM
- MAM
- Graph API
- Conditional Access
- Risk-Based CA

**Other Components**
- Federation/Dir Sync
- Internet
- Defender Advanced Threat Protection
- AIP Label Unified Client
- Endpoints
## Five categories of requirements for classification tools

<table>
<thead>
<tr>
<th>Category</th>
<th>Requirements</th>
</tr>
</thead>
</table>
| **Repositories**  | - Where can the tool detect sensitive data?  
- Does the repository type change the capabilities of the tool?  
- Can the system inspect file objects embedded in a database? |
| **Content Detection** | - What technologies are used to identify sensitive data?  
- What "out of the box" policies does the tool provide?  
- Does the tool support context-based classification, such as user or file location?  
- What language limitations does the tool have? |
| **Control**       | - Is the tool able to impose controls on the data depending on classification without other toolsets being required? |
| **Recording**     | - Does the tool provide any permanent record of the classification outcome, or is the metadata only used to implement a control at that time?  
- If metadata is persistent, what method is used? |
| **Interoperation**| - What capability does the tool have to share or consume metadata from other toolsets?  
- What capability does the tool have to "trigger" other controls or actions? |
# Data Scenarios Defined

**Repository Types Include:**

- Windows (CIFS, SMB) File Share
- Server Local Storage
- Network-Attached Storage
- UNIX File Stores
- Sharepoint
- Email Server
- Databases (SQL)
- Database (NoSQL)
- Objects and Documents Stored Within Databases
- Hadoop and Other HDFS Storage
- Enterprise Content Management (On-Premise)
- Backups
- User Endpoint (Laptops, Desktops; Including BYOD)
- User Mobile Devices (Including BYOD)
- Removable Media

**File Types Include:**

- Cloud
  - IaaS and PaaS Such As:
    - Amazon S3, EBS
    - Azure Blob, DocumentDB
  - DBaaS, PaaS and Other Middleware Solutions
  - Cloud Sharing Such As:
    - O365, Teams, OneDrive, GoogleDocs, Box
  - Cloud SaaS (SalesForce, Concur, Docusign, etc.)
  - Cloud Email Services Such As Exchange Online
Vendors do not have universal coverage of repositories
Security outcomes and the control choices

Data Classification and Information Handling Standards Requirements

- Clean Up the Data: Restrict It to Approved Locations
  - Control: Delete, Move

- Protect the Whole File: Internal Use Only
  - Control: Manage ACL, Encrypt (All Forms)

- Protect the Whole File: Internal and External Sharing
  - Control: EDRM

- Remove Parts of the File
  - Control: Masking

- Monitor and Record Activity
  - Control: Logging (SIEM)
How machine learning is increasingly involved in data classification

- Simple Text
  - Matching Keywords
  - Simple Pattern (e.g. Regex)

- Structured Documents
  - Form Detection
  - File Type

- Content-Based
  - Subject Analysis
  - Sentiment
  - Natural Language

- Image, Video and Speech
  - Object Recognition
  - Speech to Text

- Similar Content
  - Document Searching
  - Results Clustering
  - Category Creation

- Context
  - Location
  - Similarity
  - UEBA

Increasing Use of Artificial Intelligence and Machine Learning
Design Patterns for Cloud Architecture

Giordano Scuderi
ST Microelectronics – Digital Fab
Agenda

1. Introduction
2. Cloud Design Patterns
3. First architecture example
4. Second architecture example
5. Second architecture improvement using iPaaS
6. Q&A
Introduction
<table>
<thead>
<tr>
<th>Architecture style</th>
<th>Dependency management</th>
<th>Domain type</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N-tier</strong></td>
<td>Divide application into logical layers</td>
<td>Traditional business domain. Frequency of updates is low.</td>
</tr>
<tr>
<td><strong>Web-Queue-Worker</strong></td>
<td>Front and backend, decoupled by async messaging.</td>
<td>Relatively simple domain with some resource intensive tasks.</td>
</tr>
<tr>
<td><strong>Microservices</strong></td>
<td>Vertically (functionally) decomposed services that call each other through APIs.</td>
<td>Complicated domain. Frequent updates.</td>
</tr>
<tr>
<td><strong>Event-driven</strong></td>
<td>Producer/consumer. Independent view per sub-system.</td>
<td>IoT and real-time systems</td>
</tr>
<tr>
<td><strong>Big data</strong></td>
<td>Divide a huge dataset into small chunks. Parallel processing on local datasets.</td>
<td>Batch and real-time data analysis. Predictive analysis using ML.</td>
</tr>
<tr>
<td><strong>Big compute</strong></td>
<td>Data allocation to thousands of cores.</td>
<td>Compute intensive domains such as simulation.</td>
</tr>
</tbody>
</table>
Focus area in cloud architecture

- Availability
- Data Management
- Design and Implementation
- Messaging
- Management and Monitoring
- Performance and Scalability
- Resiliency
- Security
## Focus area in cloud architecture

### AVAILABILITY
- Time where system is functional and working, measured as a percentage of the uptime.
- **Applications must be designed to maximize availability.**

### DATA MANAGEMENT
- Data is typically hosted in different locations and across multiple servers
  - data consistency must be maintained, data will typically need to be synchronized across different locations.
- key element of cloud applications, influences most of the quality attributes.

### DESIGN AND IMPLEMENTATION
- Decisions made during the design and implementation phase have a huge impact on the quality and the total cost of ownership of cloud hosted applications and services.

### MESSAGING
- Applications requires a messaging infrastructure that connects the components and services, ideally in a loosely coupled manner in order to maximize scalability.
- Challenges: ordering of messages, idempotency, ...
### Focus area in cloud architecture

<table>
<thead>
<tr>
<th>MANAGEMENT &amp; MONITORING</th>
<th>RESILIENCY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Applications should help exposing information operations team can use to manage and monitor the system</td>
<td>Resiliency is the ability of a system to gracefully handle and recover from failures.</td>
</tr>
<tr>
<td>Applications should support changing business requirements and customization without requiring the application to be completely stopped</td>
<td>Due to the nature of cloud hosting there is an increased likelihood of both transient and more permanent failures.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PERFORMANCE AND SCALABILITY</th>
<th>SECURITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance is an indication of the responsiveness of a system to execute any action within a given time interval</td>
<td>Capability of a system to prevent malicious or accidental actions outside of the designed usage, and prevent disclosure or loss of information.</td>
</tr>
<tr>
<td>Scalability is the ability of a system to handle increases in load without impact on performance</td>
<td>Applications must be designed and deployed in a way that protects them from malicious attacks, restricts access to only approved users, and protects sensitive data.</td>
</tr>
</tbody>
</table>
Cloud design patterns

• Patterns are a widely used concept in computer science
  • Describe **good solutions to recurring problems in an abstract form**.
  • Can be applied regardless of used technologies

• Let’s see some patterns as example…

• References:
  • [https://www.cloudcomputingpatterns.org](https://www.cloudcomputingpatterns.org)
  • [https://docs.microsoft.com/en-us/azure/architecture/](https://docs.microsoft.com/en-us/azure/architecture/)
The distributed nature of cloud applications requires a messaging infrastructure that connects the components and services, ideally in a loosely coupled manner in order to maximize scalability.
Publisher-Subscriber pattern

- Enable an application to announce events to multiple interested consumers asynchronously, without coupling the senders to the receivers
Scalability is the ability of a system to handle increases in load without impact on performance.
Cache-Aside Pattern

- Load data on demand into a cache from a data store
- Improve performance
- Helps to maintain consistency between data held in the cache and data in the underlying data store
Static Content Hosting Pattern

- Minimizes web hosting compute costs
  - Especially so for web sites that consist only of static content
- Improves end-user content performance
  - Serving static content from a CDN (Content Delivery Network) can save on compute and memory utilization of web servers
Performance is an indication of the responsiveness of a system to execute any action within a given time interval.
• Control the consumptions of resources used by a single account, or even an entire application which is using the service
Design and Implementation patterns
Ambassador pattern

- Useful for off-loading common tasks such as monitoring, logging, routing, and resiliency patterns in a language agnostic way.
Pipes and Filters pattern

- Decompose complex processing tasks into a series of separate elements that can be reused
Resiliency is the ability of a system to gracefully handle and recover from failures.
Circuit Breaker Pattern

Closed
- If operation succeeds
  return `result`
- Else
  increment failure counter
  return `failure`

Success threshold count reached
(Reset failure counter)

Failure threshold reached
(Start timeout timer)

Half-Open
- If operation succeeds
  increment success counter
  return `result`
- Else
  return `failure`

Timeout timer expired
(reset success counter)

Open
- return `failure`

Operation failed
(start timeout timer)
The Retry Pattern

- The Retry pattern is designed to handle temporary failures (Transient Errors).
- Failures are assumed to be transient until they exceed the retry policy.
Queue-Based Load Leveling Pattern

- Use a queue to act as a “buffer” between requestor generators and request services
- Queue decouples the tasks from the service
  - Services can work at their own pace
Security is the capability of a system to prevent malicious or accidental actions outside of the designed usage, and to prevent disclosure or loss of information.
• solve the problem of controlling access to a data store where the store can't manage authentication and authorization of clients.
Gatekeeper pattern

- Provide an additional layer of security, and limit the attack surface of the system
Data Patterns
Sharding Pattern - Lookup

- Divide a data store into a set of horizontal partitions or shards. This can improve scalability when storing and accessing large volumes of data.

**Sharding logic:**
- Route requests for customer 5 to Shard A
- Route requests for customer 21 to Shard C
- Route requests for customer N to Shard X

Diagram:
- Application Instance
  - Query: Find info on Customer 5
  - Shard A
  - Shard B
  - Shard C
  - ...
Architecture 1: serverless web app
The example architecture describes a generic single-page application.

YOU CAN DEPLOY THIS ARCHITECTURE:
https://github.com/mspnp/serverless-reference-implementation/blob/v0.1.0-update/README.md
### Example 1: architecture considerations

#### SCALABILITY
- **Functions.** Functions scale based on the traffic (unless fixed plans are used). There is a limit to the number of concurrent function instances, but each instance can process more than one request at a time.
- **API Management.** API Management can scale out and supports rule-based autoscaling. The scaling process takes at least 20 minutes which must be considered depending on which kind of traffic is expected in your application.
- **Cosmos DB.** Use partitioning to scale individual containers in a database to meet the performance needs of the application. Select the best partition key is an important decision that will affect the application performance.

#### SECURITY
- **HTTPS** should be enforced throughout all the components to improve security.
- Re-use standard services or libraries to implement the **authentication**.
- To secure the back-end components, backend services should **restrict access** to the API service only, this reduces the attack surface of the application.

#### OPERATIONAL
- **Separate resources** for production, development, and test environments in dedicated and segregated area to allow a correct assignment or access rights.
- Implement **monitoring and logging capabilities** in your app and configure the logging, monitoring and alerting of the services you are using.
- **Deploy using automation** (IaC) to simplify the operational aspects and avoid human errors when deploying in production.

#### BUSINESS CONTINUITY
- **Functions.** Functions scales based on the traffic (unless fixed plans are used). But there is a limit to the number of concurrent function instances
- **API Management.** API Management can scale out and supports rule-based autoscaling. The scaling process takes at least 20 minutes which must be considered depending on which kind of traffic is expected in your application.
- **Cosmos DB.** Use partitioning to scale individual containers in a database to meet the performance needs of the application. Select the best partition key is an important decision that will affect the application performance.
Architecture 2: Advanced analytics
This architecture allows you to combine data at different scale, and to build custom machine learning models.
Example 2: architecture considerations

<table>
<thead>
<tr>
<th>SCALABILITY</th>
<th>SECURITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>• ?</td>
<td>• ?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>OPERATIONAL</th>
<th>BUSINESS CONTINUITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>• ?</td>
<td>• ?</td>
</tr>
</tbody>
</table>
Use Case:
Decouple ingestion layer by leveraging an iPaaS

Flow is:
1. iPaaS Engine retrieve the credentials to connect to the source and destination
2. Credentials are used to authenticate to Object Storage and source system
3. iPaaS Engine manage the data transfer from source system to destination (data masking might be done in this stage before writing data into destination)
Q&A
Thank you