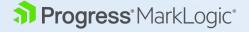# MarkLogic Server

Key Concepts in this Unit

- What is MarkLogic Server?

- How top companies use MarkLogic Server?

- Getting started.

What is MarkLogic Server?

"A platform-independent, enterprise-grade, schema-agnostic NoSQL document database"

NoSQL Document database

- Supports the following type of data:

  - Structured: XML and JSON

  - Unstructured: Text

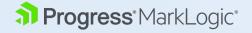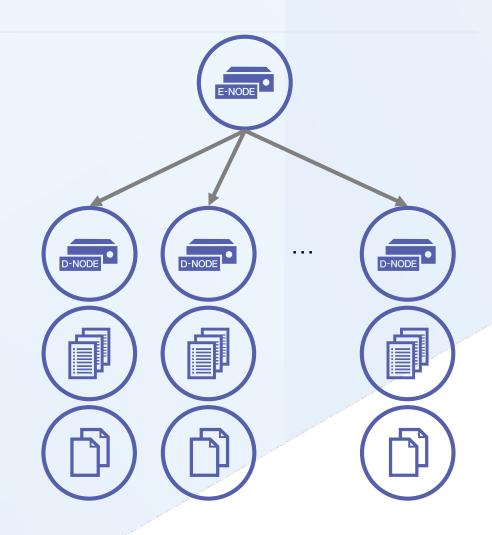 - Graph: RDF Triples

  - Binary: PDF, Audio, Video, etc.

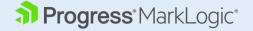# NoSQL Document database

- Horizontal scaling using commodity hardware.

  - No master or slave hosts.

- Distributed query resolution.

  - Each Data Node manages its own set of documents and Indexes.

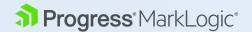- MarkLogic supports querying with SQL-92.

Schema-agnostic

- Load data "as-is"

  - No need to define a super-schema prior to loading your data.

- Schema-**agnostic** does NOT mean schema-*less:*

  - Document structure is tracked along with content.

  - Enforce structure and datatype if needed.

```xml
<profile>
  <name>
    <given>Jane</given>
    <surname>Roe</surname>
  </name>
  <dob>1981-10-22</dob>
  <email>jane.roe@example.com</email>
  ...
</profile>
```

```json
{
  'user' : {
    'first' : 'John',
    'last' : 'Doe',
    'dob' : '1979-06-12',
    'email' : [
      'jdoe@acme.com',
      'john@domain.com'
    ],
    ...
  }
}
```
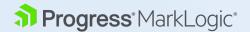
Enterprise-grade

- ACID compliance via Multi-Version Concurrency Control (MVCC)

- High Availability – Minimizes, if not eliminates, downtime.

- Database Replication – Minimizes data loss in the event of a major disaster.

- Backups – Scheduling.

- Recovery – Supports point-in-time.

- Encryption-at-rest – Industry standard support.

- Encrypted communication.

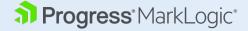- Rolling upgrades.

Supported Platforms – v11.3

- <u>Operating Systems</u>:

    - RHEL 8/Rocky Linux 8

    - Mac OSX* - 10.14+

    - Microsoft Windows – x64 - 10*, Server (2012, 2016, 2019)

- Cloud:

    - <u>AWS</u>

    - <u>Azure</u>

- Virtualization/Containers

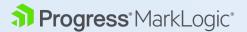   - <u>Docker</u>, <u>Kubernetes</u>

* Development purpose only

MarkLogic Database is NOT…

- Relational Database Management System (RDBMS)

  - MarkLogic Server supports SQL92 other server specific functions.

  - Content is stored as documents (multi-nested) and/or graphs, not rows (flat).

- Data Lake

  - Content is indexed and searchable as soon as it's loaded.

- Web Container

  - HTTP App Server that can support Single Page Applications (SPA) like React and VueJS, but not Servlets, PHP, Python, etc.

MarkLogic Database is NOT…

- "Just a document database"

  - Pipelines can be triggered based on document structure, collection or Universal Resource Identifier (URI).

  - Stored Procedures can be created that execute within the MarkLogic cluster using Data Services or REST API extensions.

  - Perform custom actions (non-pipeline in nature) on document events (create, update, delete) before or after the source transaction commits.

  - Track the state of documents between when it was valid in the real world (valid time) and when the document got stored (system time) using bi-temporal feature.

- An Optical Character Recognition (OCR) Tool

  - The "Converters" library that allows for content extraction from binary documents.

  - Multimedia files can only extract metadata, such as, resolution, image size, codec, etc.

Power Packed Features

SEARCH

SEMANTIC INFERENCE

TELEMETRY

TEMPLATE DRIVEN EXTRACTION

TIERED STORAGE

MASKING & REDACTION

MONITORING & MANAGEMENT

ELASTIC LOAD BALANCER SUPPORT

EXTERNAL KEY MANAGEMENT

GEOSPATIAL

HADOOP DISTRIBUTION

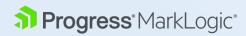BITEMPORAL

COMPARTMENT SECURITY

MANAGEMENT API

BI TOOLS INTERFACE

ADDITIONAL ENHANCEMENTS

Progress® MarkLogic®

Lab: Access eBooks and White Paper

Customer Use Cases

Progress® MarkLogic®

Development flexibility

- DTIC has more than 30 data sources consisting of unstructured and structured data, and needed a way to make that information discoverable, shareable, enriched and secure. The organization is using a MarkLogic Data Hub to achieve these goals.

> "MarkLogic ingested 4 data sets, implemented search capabilities and element-level security, and demonstrated semantic capabilities—all in six weeks with two developers."
>
> **ANDREW PEDRICK**
> INFORMATION ARCHITECT, DTIC

Centralized from silos

- Amgen faced numerous data silos, which included different types of data as well as ownership of that data. Additionally, definitions of the data varied, such as sodium bicarbonate and NaHCO3. Amgen is using a MarkLogic Data Hub on AWS to ingest, harmonize and link data together for a 360 view of information, improving drug delivery technology, product portfolio management, logistics, translational sciences, and reducing risk.

"We've derived operational benefits in terms of cost reduction, efficiencies, and, in analytics, we've come up with better reporting mechanisms, which helps in risk management."

DR. ALICE "CLARE" AUGUSTINE
TAXONOMY MANAGEMENT LEAD, AMGEN

Fast and Scalable

- Legacy system slowed productivity, innovation, and content delivery
- Growing content required massive scalability and faster response times

> **...what we got in 20 seconds on SQL took us 200 milliseconds in NoSQL...."**
>
> "Our legacy system took up to 30 minutes or an hour sometimes [to publish a video clip]. We compared the MarkLogic NoSQL technology to some SQL vendors, and what we got in 20 seconds on SQL took us 200 milliseconds in NoSQL — orders of magnitude faster. So we said let's move iPlayer to this."
>
> **ALLAN DONALD**
> SENIOR PRODUCT MANAGER FOR EDITORIAL METADATA, BBC

See more use cases…

- Check out other use cases for [different industries](#)…

ABN·AMRO

ABN Amro Brings Vast Amounts of Data into Central Operational Trade Store

VIEW CASE STUDY

SONY

Sony Integrates Data to Improve Global Sales to Finance Operations for its $9B Business

VIEW CASE STUDY

SUFFOLK

Suffolk Construction Uses MarkLogic to Leverage Data, Helping Customers Build Better, Faster and Cheaper than Before

VIEW CASE STUDY

NBC

NBCUniversal Harnesses 40 Years of Comedy from SNL

VIEW CASE STUDY

HealthCare.gov

CMS Speeds Time-to-Production, Creates Seamless Online Shopping Experience

VIEW CASE STUDY

CHHS
California Health & Human Services Agency

California Health & Human Services (CHHS) Breaks Down Data Silos to Better Serve Over 26M People
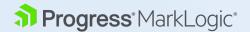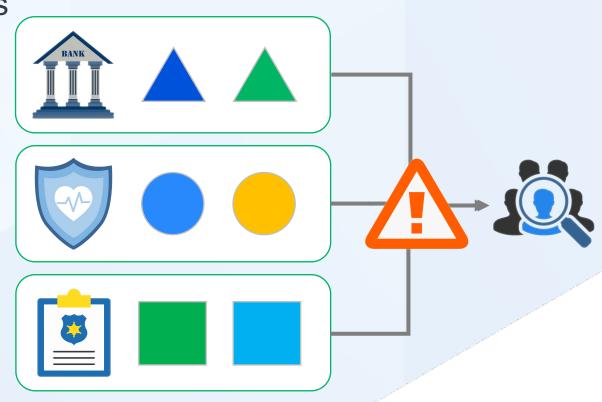
VIEW CASE STUDY

Progress® MarkLogic®

Common Large Data Problems

- Related information is scattered across different silos.

- Information is structured differently across silos.

  - A side-effect as they are structured based on system-specific needs/design.

- Integrated search across all silos is difficult, if not impossible.

How MarkLogic Solves This

- Load data "as-is", no pre-defined schema needed.

  - Data is immediately indexed and searchable.

- Curate based on a need/feature.

  - An iterative process of transforming or generating the desired shape of data.

- Serve curated data downstream

  - REST API, Data Services

  - ODBC Connector
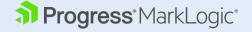
  - Graph

Getting started

Resource requirements

- For the purposes of this course/development:
    - CPU: 2 cores, more is better
    - RAM: 4 GB RAM, more is better
    - Disk: 10GB of disk space
        - Prepared environment image would require 20 GB
- Sizing requires a representative data set to be loaded into MarkLogic, have the corresponding indexes configured, etc. before any recommendation could be made.

Download and Installation

- Installer can be downloaded from developer.marklogic.com.

  - There is NO evaluation version.

  - Licensed features like language are already included and activated upon license configuration.

- "Converters and Filters" are only required if you intend to extract information from binaries.

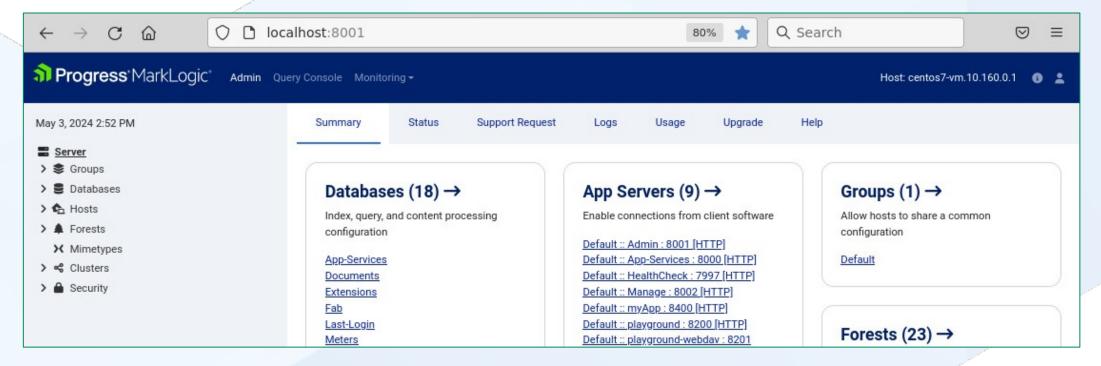- Installation procedures are available at docs.marklogic.com or progress.com/marklogic/documentation .

Progress® MarkLogic®

Licensing

- Developer License is free and enables all features.

  - Limited to 1TB of data

  - Valid for 6 months, renewable on demand.

  - Non-commercial use

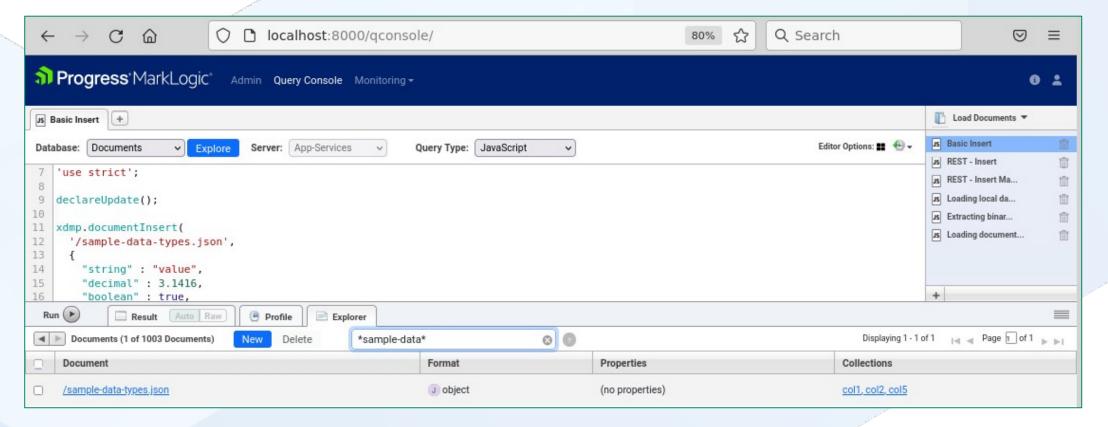- For production license, please contact sales@marklogic.com
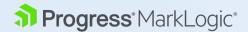
Built-in interfaces
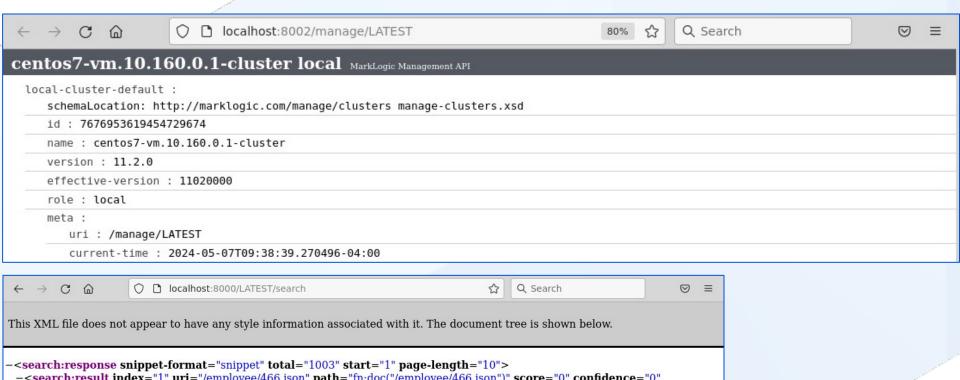
- Administrative (Admin) UI

Built-in interfaces

- Query Console

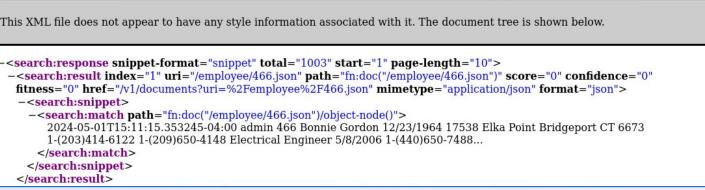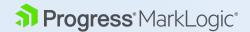# Built-in interfaces

- REST API

Progress® MarkLogic®

Lab: Install, Initialize and Access MarkLogic

Recap

- MarkLogic Server is a platform-independent, enterprise-grade, schema-agnostic NoSQL document database.

- MarkLogic Server is used by companies of different industries across the world.

- Free Developer License allows for up to 1TB of data across any number of hosts.