10TH EXPERT GROUP MEETING ON

Statistical Data and Metadata eXchange

JANUARY 25-28, 2021

SDMX in a Big Data Architecture

January 27, 2021

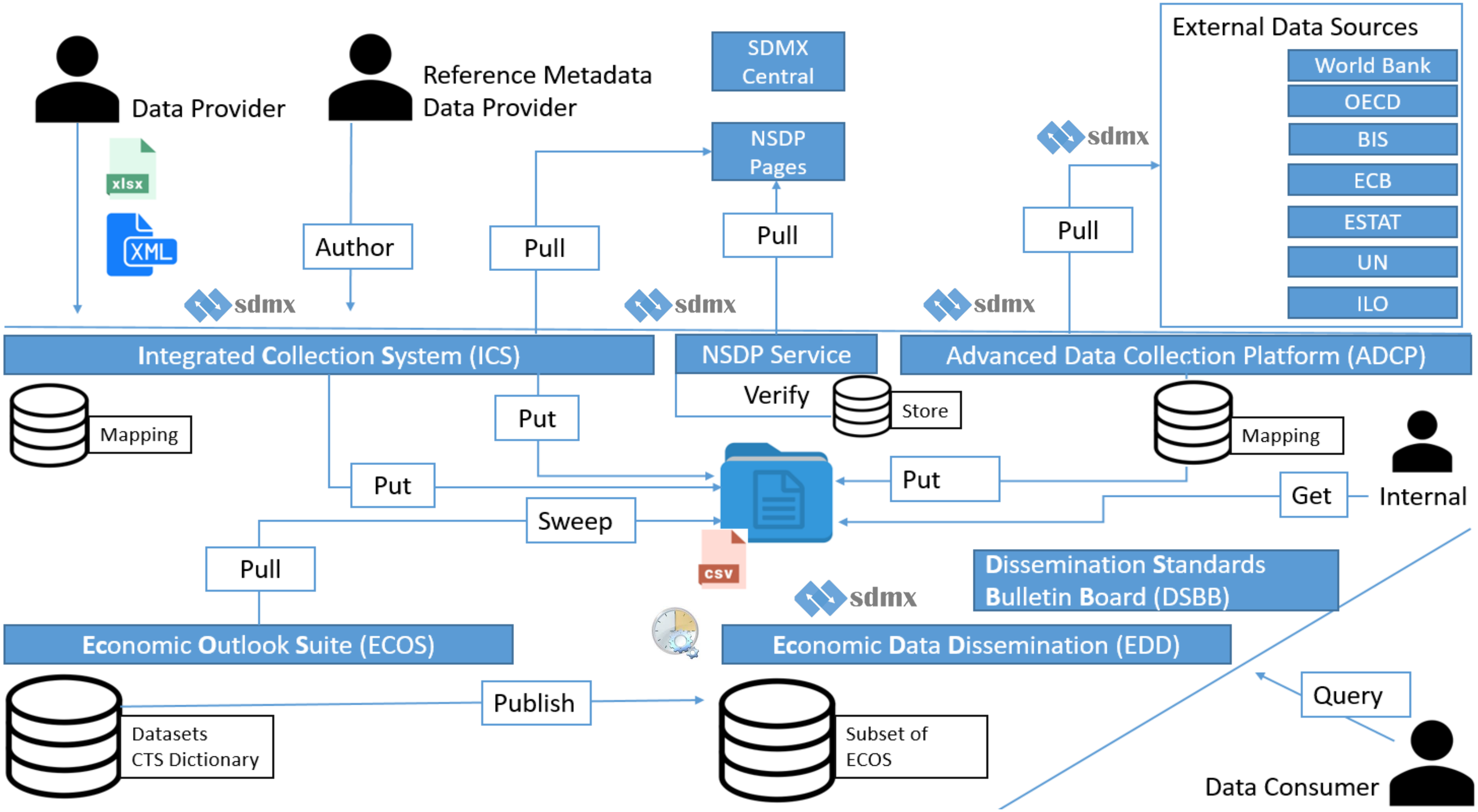
José Deodoro & Ryoichi Yakamoshi
Data Collection Platform – IMF

The views expressed in this presentation are those of the authors

Outline

- SDMX architecture in IMF
- Issues and Challenges
- Target characteristics
- Proposed Architecture
- Evaluation
- Next Steps

Present SDMX architecture



sdmx
Statistical Data and Metadata eXchange

SDMX large file issue

Dataset	Provider	Download	Processing	Input	Output
Trade in Value Added (TiVA): December 2016	*OECD*	00:38:14	00:32:30	2.98 GB	7.19GB
Trade in Value Added (TiVA): Origin of Value Added in final demand	*OECD*	00:26:50	00:23:50	2.26 GB	5.91 GB
Trade in Value Added (TiVA): Principal indicators	*OECD*	00:22:24	00:18:15	1.78 GB	3.95 GB
EU trade since 1999 by HS2,4,6 and CN8 - daily updated	*Eurostat*	00:16:13	00:17:12	1.48 GB	2.32 GB

Related challenges

- Scheduler may interrupt long processes
- Bandwidth is affected during batch process
- Data Validation errors affect re-batch process

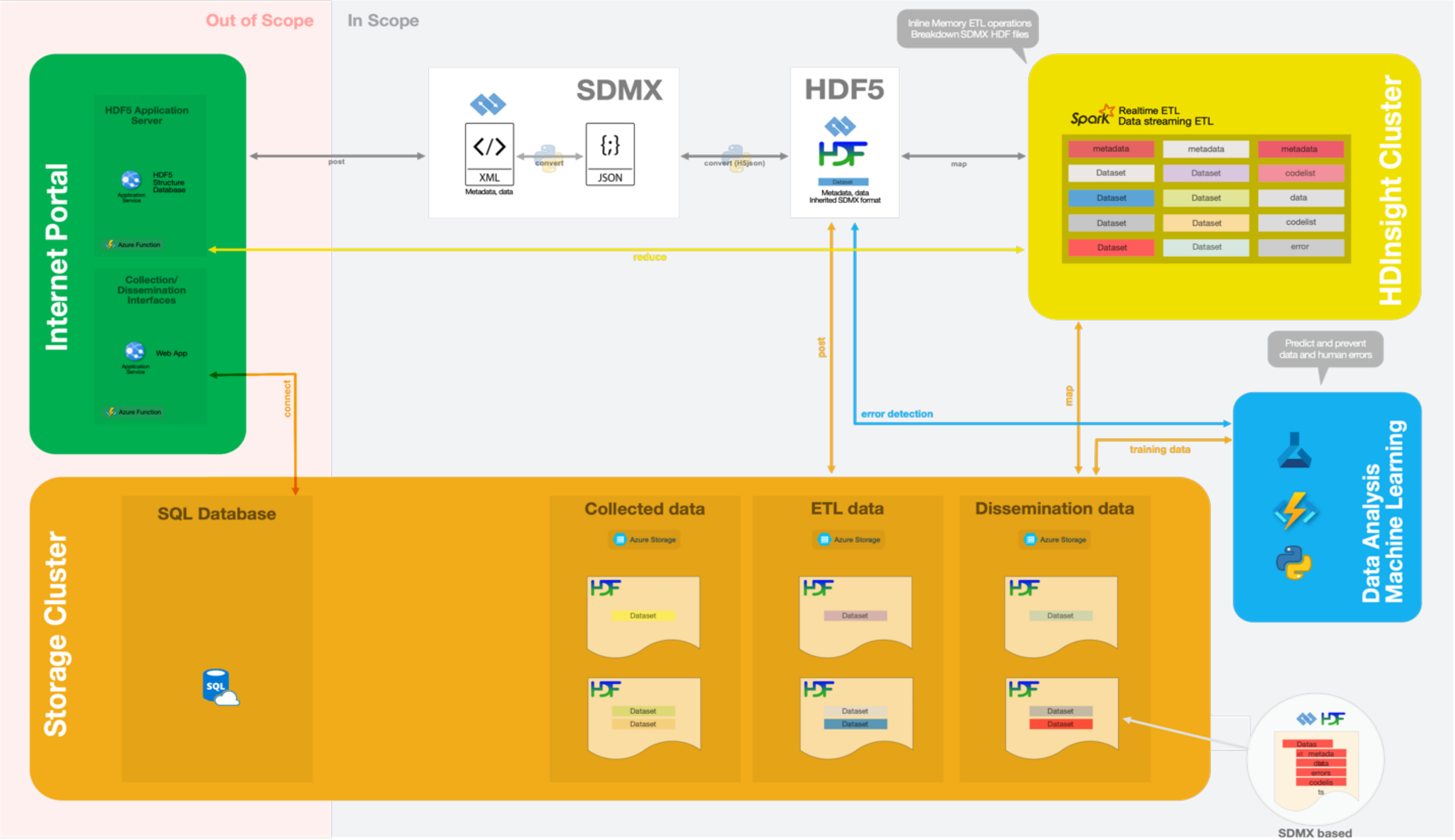
Desired characteristics

- Optimized I/O operations
  by squeezing data size
- In-memory processing
  improved performance, reduced bandwidth, shorter batches
- Pipeline for Machine Learning (ML)
  reducing data format error using prediction and validation

Proposed Components

- HDF5
  - Structure based (Tree architecture)
  - Binary
  - Applied to data science
- Using Spark (Databricks)
  - Cloud environment
  - In-line memory processing
  - Real time (streaming) process
- Machine Learning (ML)
  - Data Validation (Prediction and Privation)
  - Real time Machine Learning

Proposed Architecture



HDF5 Simple Data flow and Architecture

HDF5 Evaluation

Source	SDMX file size (MB)	HDF5 file size (MB)	Compression rate	SDMX I/O (sec)	HDF5 I/O (sec)
WEO_PUB_APR2020	6.2	2.4	38%	0.5	0.1
WEO_PUB_APR2020_Quarterly	8.1	3.2	39%	0.5	0.12
WEBApr2020	100.6	33.4	33%	1.5	0.5
WEO_POB_OCT2019	25	9.8	39%	1.2	0.2

Projected performance

Dataset	Provider	Download	Processing	Input	Output	Est. Size
Trade in Value Added (TiVA): December 2016	*OECD*	00:38:14	00:32:30	2.98 GB	7.19GB	1.2 GB
Trade in Value Added (TiVA): Origin of Value Added in final demand	*OECD*	00:26:50	00:23:50	2.26 GB	5.91 GB	.9 GB
Trade in Value Added (TiVA): Principal indicators	*OECD*	00:22:24	00:18:15	1.78 GB	3.95 GB	.7 GB
EU trade since 1999 by HS2,4,6 and CN8 - daily updated	*Eurostat*	00:16:13	00:17:12	1.48 GB	2.32 GB	.6 GB

Next Steps

- Proof of concept under discussion at the Fund

Thank you

Jose Deodoro
jdeodoro@imf.org

Ryoichi Yamakoshi
ryamakoshi@imf.org

sdmx
Statistical Data and Metadata eXchange