Irving Fisher Committee on
Central Bank Statistics

12th biennial IFC Conference: "Statistics and beyond: new data for decision making in central banks"

22-23 August 2024

Implementing Statistical Data Metadata Exchange (SDMX) in Centralised Information Management System (CIMS) – challenges, solution design and benefits[1]

Debasis Nandi and Prakash Madda,
Reserve Bank of India

[1] This contribution was prepared for the conference. The views expressed in this publication are those of the authors and do not necessarily represent the official views of the Committee, its members, or the BIS.

Implementing Statistical Data Metadata Exchange (SDMX) in Centralised Information Management System – Challenges, Solution Design and Benefits

Debasis Nandi, and Prakash Madda[1]

Abstract

Implementation of an SDMX-based data architecture in the Reserve Bank of India necessitated innovative solutions for managing dataflow through templates from over 10,000 regulated entities, surveys, and other systems. Data structure definitions (DSDs) and related SDMX artefacts / metadata facilitate automated system-based data transmission and data governance with all Regtech and Suptech[2] features. Data pipelines and analytical tools are built and innovative SDMX converter generates time series data from regular filing of statements. Collection and processing of micro-data and unstructured data, generic application for SDMX analytical data query and visualisation support data-driven policies as well as other analytical and dissemination needs.

Keywords: Data Warehouse, OLAP, Big Data, CIMS, SDMX, RegTech, SupTech

[1]Shri Debasis Nandi (dnandi@rbi.org.in) is Director and Shri Prakash Madda (prakashm@rbi.org.in) is Assistant Adviser in the Department of Statistics and Information Management, Reserve Bank of India. The views expressed in this paper are of the authors and not the institution to which they belong.

[2] **RegTech** is a platform that enhances regulatory compliance through automation, making the process more efficient and cost-effective. It leverages technology to deliver regulatory requirements more efficiently and effectively. **SupTech** uses technology to enhance regulatory supervision by enabling efficient data collection, analysis, and decision-making.

I. Introduction

1.1 Data are the essential input for central banks' decision-making process in their core functions of monetary policy, financial stability, supervision of market and payment systems, and conducting supervisory and regulatory oversight of regulated entities. Accordingly, central banks persistently work to ensure that their data management systems cover all the domains of activities and they get data of acceptable quality within a reasonable time frame, where analytical capability, both in terms of economic/statistical analysis and the technology tools for enabling such analysis, are on high priority. Data warehousing and Online Analytical Processing (OLAP) enable quick analysis of multidimensional data sets.

1.2 The Reserve Bank of India (RBI) was one of the early adopters of data warehousing (DW) among central banks in 2002 with an objective to collect, compile and disseminate wide range of data collected from regulated entities, monetary policy surveys, various government agencies and its internal systems. This initiative resulted in the creation of the Database of Indian Economy (DBIE) public web portal, which serves as a comprehensive resource for a wide range of information related to the Indian economy. DBIE contains data relating to diverse areas such as financial markets, banking indicators, macroeconomic aggregates, external sector data, etc., and facilitate policymakers, researchers, analysts, and the public with access to timely and reliable data.

1.3 With explosion in data volume, traditional information management systems face challenges of maintaining efficiency and of catering to the need of internal and external stakeholders. Availability and applicability of high frequency large volume and unconventional textual data (popularly known as Bigdata) have added another major dimension to it. Data requirements of business users are becoming more dynamic, which leads to frequent changes in data structure. The traditional information management system built on DW technology with Relational Database Management System (RDBMS) data architecture is not able to meet such novel requirements.

1.4 The surge in data volume, variety, and velocity in new modern world has led to the development of new information management technologies which leverage distributed data storage techniques and parallel processing methods for efficient handling of large-scale data. Accordingly, new technologies, tools, and frameworks based on Hadoop Distributed File System (HDFS) is specifically designed for managing and analyzing large volumes of data (Shvachko *et al.* 2010). MapReduce is a programming model and processing framework designed for parallel processing of large datasets across distributed computing clusters (Zaharia *et al.* 2010). Apache Hive and Apache Pig provide high-level abstractions and query languages for processing and analyzing large datasets (George 2011 and Olston *et al.* 2008). Apache Spark, as outlined by White (2009), offer in-memory data processing capabilities and support for a wide range of data processing tasks, including batch processing, stream processing, machine learning, and graph processing. These technologies are appropriate to address most of the information management challenges faced by central banks.

1.5 This paper provides an overview of the RBI's new data lake solution and implementation of SDMX project. It covers the adoption of SDMX, and development of various tools and technologies used for data collection, compilation methods, dissemination strategies, technological advancements, and alignment with Suptech and Regtech solutions.

1.6 The rest of the paper is organized as follows: Section II focus on evolution of Information management at the RBI, Section III gives an overview of the Centralized Information Management System (CIMS) set up by the RBI, and Section IV elaborates Statistical Data and Metadata Exchange (SDMX) solution architecture in CIMS. Implementation challenges for

SDMX systems are discussed in Section V and key takeaways are described in the concluding Section VI.

II. Evolution of Information management at Reserve Bank of India

2.1 The RBI has a long history of collecting data from regulated entities for its operational needs. The scope of data collection has been consistently rising to meet the growing information needs for both business purposes as well as for planning and policy making. Advent and increased availability of computer technology and computational power led to their adoption by the RBI and later, gradual introduction in the banking industry, which gathered pace with increased automation of banking business and provided appropriate atmosphere to move towards collection of data in electronic formats.

2.2 In 2002, the Reserve Bank of India (RBI) set up the Database on Indian Economy (DBIE), an enterprise-wide Data Warehouse (DW), with the broad objectives of (a) developing an integrated repository of current and historical data encompassing all operational and research areas; (b) building a state-of-the-art decision support infrastructure with OLAP and dashboard capabilities; and (c) creating a user specific data mart based on central repository. Initially, the DW was made operational for web-based access over the closed user group network for users across offices of the RBI. In 2003, the publishable part of the data was made available to the public and it was extended from a data repository to a platform for (a) data collection and validation; (b) data processing; and (c) data dissemination and publication.

2.3 The RBI started implementing eXtensible Business Reporting Language (XBRL) technology for collecting data from regulated entities on banking returns during 2008. Out of nearly 250 regular returns / statements flowing to the RBI, over 150 were gradually covered in the DW System by end-2022. Majority of them were collected through XBRL format. This was accompanied by dismantling silo information systems and moving towards centralizing the information processing in the DW.

2.4 As the DBIE matured and became an important source of data dissemination, including for multilateral agencies, RBI started bringing out its important regular statistical publications through the DBIE.

2.5 With growing demand to integrate micro-level data from different internal units into the DBIE, technological obsolescence and scalability limitations became pressing issues. Therefore, the RBI decided to overhaul its data collection and compilation system by implementing Data Lake (big-data Hadoop technology) and SDMX to exploit the full potential of the emerging technologies. Data lake technology facilitates efficient collection, storage and processing of high volume and unstructured data. Implementation of SDMX enables management of data life cycle for each data element and facilitate putting in place system driven data governance framework.

III. Centralised Information Management System (CIMS)– The Big Shift

3.1 CIMS has been designed as a comprehensive database management system for complete lifecycle management of data, covering data collection, processing, storage, standardization of reporting, dissemination, and data analytics. Its major features are: (I) automation of data transmission from source entities (banks / other regulated entities); (ii) development of state-of-the-art SDMX data architecture; (iii) management of high volume structured and unstructured data using Bigdata open-source technology like Hadoop; and (iv) development of integrated data analytics, visualization, and data dissemination.

3.2 Functionalities of CIMS:

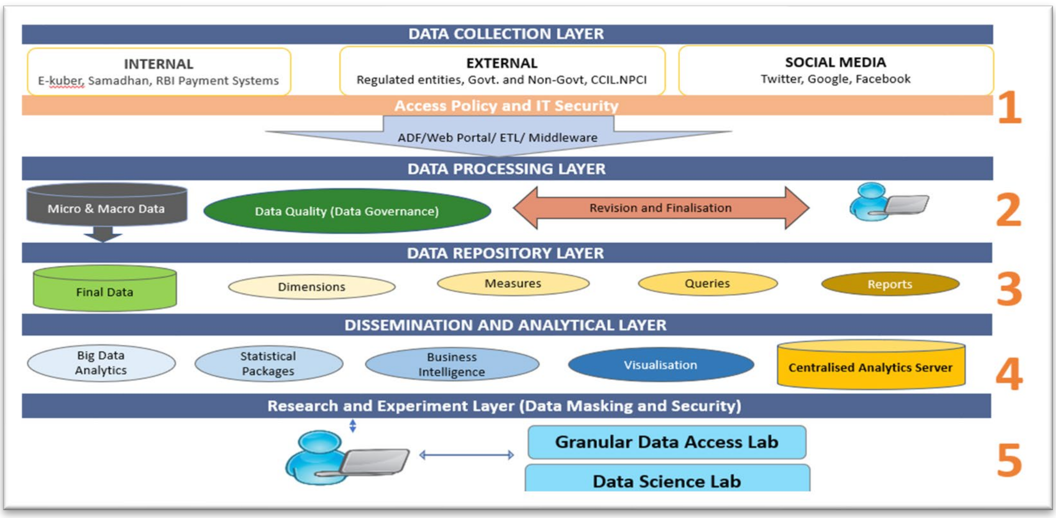
The main objective of CIMS is to support data driven policy making and monitoring by providing quality data and analytical platform for cross domain analysis and data dissemination to public. Some of important features are:

- Centralisation of various types of data under one umbrella which includes financial statements, regulatory reports, compliance information, and other relevant data reported by over 10,000 regulated entities,
- Multiple channels for data sharing such as a) seamless data collection using server-to-server interface; ii) file upload at web portal; and iii) web forms for small sized financial intermediaries,
- Supports RBI's supervisory activities by providing a platform for monitoring and analyzing regulatory compliance, risk exposures, and other key indicators of financial institutions.
- Unstructured data collection and text analytics using web scrapping, social media, and news.
- Integrated analytical platform to users to carryout visual and advanced statistical analysis using R and Python programming. This helps cross domain analytics in identifying emerging risks, trends, and areas requiring regulatory intervention.
- Automation in processing data and generating various publications for public data dissemination.
- Flexible and scalable Hadoop based data repository covering all information management needs with safe and secure data governance/ data quality functionalities.
- Data governance web-portal for creation and maintenance of metadata and monitoring data submission timeline, data validation and data quality parameters.

3.3 Conceptual Architecture

CIMS is structured into five layers, each serving a specific purpose within the system.

Picture-1: CIMS conceptual Architecture



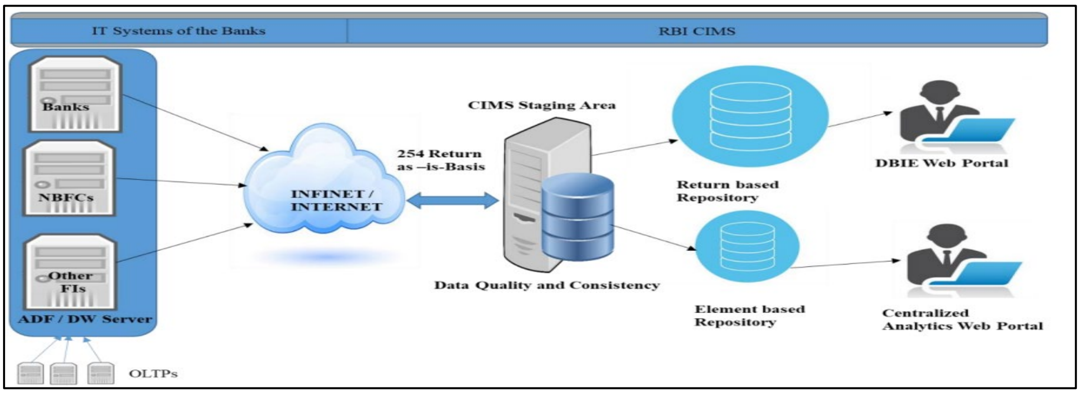
3.4 Data collection layer:

This layer facilitates ingesting data from various sources, including banks, financial institutions, regulatory filings, external data collected from Government of India and other agencies, and granular transaction data (especially payment systems and government securities) from internal IT systems. RBI collects large amount of data from regulated entities for its regulatory and supervisory needs[3]. It also conducts various surveys to gather specific information on various aspects of the economy and financial systems from households and enterprises, as policy inputs. A substantial chunk of financial sector data are generated as by-product of transaction systems.

3.5 Implementation of Automatic Data Flow (ADF):

3.5.1 In India, the top 50 major banks account for over 90 per cent of banking business. These banks were requested to generate the data required for submission to RBI in prescribed formats using automated scripts and place the same at an identified location(server) for onward transmission to the RBI. A specific software component was developed for server-to-server automated data collection using middleware technology. These major banks are now connected using automated data collection for reporting data in the current format.

Picture-2: ADF architecture



3.5.2 Medium and small-sized intermediaries may not find it cost effective to set up an elaborate infrastructure for automated data transfer. Such users upload the required data in a specific format through a portal. Smaller entities can also use the screen template in a dedicated data collection portal for their data submission.

3.5.3 Data from internet sources (structured and unstructured) including various international/domestic official agencies, social media, newspapers, and ecommerce sites, are collected by using web-scrapping tools.

3.6 Data processing Layer: This layer mainly focusses on validation of data formats, data contents validations, and internal consistency among various elements of data. The data processing layer involves transforming raw data into a format that is suitable for reporting and loading into the database for use in further analysis. This transformation may include data cleansing, normalization, aggregation, and enrichment to ensure data accuracy, consistency, and completeness. The system has facility to send error logs with pinpointing

[3] https://www.rbi.org.in/scripts/BS_Listofallreturns.aspx

errors through automated system-driven e-mails and the same files made available in the portal for necessary download.

3.7 Data Repository Layer: This layer provides a comprehensive data and metadata repository to store all data. There is separate repository built for storage of return-based data named as Return Based Repository (RBR)" and SDMX based data named as "Element Based Repository (EBR)" with interoperability using mapping tables and an innovative converter application.

3.8 Dissemination and Analytical Layer: It is the data presentation layer, which serves as the interface based on access right management module through which users interact with the data and analytical tools. It encompasses user interfaces, dashboards, visualization tools, and reporting capabilities that enable users to access and visualize data, generate reports, and make informed decisions based on the insights derived. This layer has multiple portals developed for specific user classes, such as, internal users, users from regulated entities and public. A web portal named as Data Base of Indian Economy (DBIE)[4] for data access by external users to access reports/ interactive dashboards/ other facilities which are publicly available and accessible without logging in the CIMS system.

3.9 Research and Experimentation layer: This layer includes tools and algorithms for performing various types of data analysis, such as descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics. These analytics techniques help uncover patterns, trends, correlations, and insights hidden within the data by the inhouse Data scientist. Additionally, a secure user login based Granular Data Access Lab (GDAL) is under development to facilitate granular level data access, which is not publicly available, to external users in restricted environment for data analysis without compromising the privacy and confidentiality.

IV. SDMX Solution Architecture

4.1 Being a full-service central bank, the activities of the RBI span widely and so is the genesis of complexity of its database. Different departments collect data on multiple variables, which are often manifestation of related concepts, but differ on account of institutional or geographical coverage, inclusion or exclusion of certain sub-items arising from policy prescriptions, and reference dates for reporting on account of statutory or legacy reasons. The data are also required on distributional aspects by economic sectors, institutional sectors, maturity profile, geographical locations, or demographic characteristics of the parties, to their identities, particularly where the exposures are large. Such diversity in data characteristics and their diverse usage across departments often result in consistency issues with each other, when measured through the traditional reporting structure, which was an important challenge in modelling data in the CIMS, while replacing the earlier data warehouse.

4.2 A series of steps were taken to harmonise the concepts and definitions and rationalise the reporting system, the impact of which was significant in reducing inconsistencies. Nevertheless, a more concrete and comprehensive strategy was necessary to overhaul the entire value chain of data, from collection, modelling, storage, retrieval and usage by the end-user. With many flavours of similar variables available in the system, it was necessary that the data are accompanied by well-structured and comprehensive metadata to fully describe the data and facilitates processing and usage.

[4] https://data.rbi.org.in/

4.3 An international standard facilitating the exchange of statistical data and metadata, in the form of Statistical Data and Metadata Exchange (SDMX) was an appropriate fit for the requirement of the Reserve Bank. It establishes common formats, protocols, and guidelines for sharing data among statistical agencies, international organizations etc. in seamless manner. SDMX key components include DSD for data format, XML-based Message Format for data exchange, Web Services-API for automated integration, Code Lists for standardization, Metadata for contextual information, and Registry for central repository and maintenance tool for all of the structures including Code lists, Concepts and Data Structures.

4.4 SDMX Design Approach:

4.4.1 The SDMX ecosystem was designed, keeping in view seamless management of the entire lifecycle of data starting its collection from the regulated institutions to its consumption by the internal users and dissemination to public, where applicable. The central part of the design is the data repository for data elements with all the attributes and their inter-relationships. The main objective was to transform the data collection from purpose specific to data specific templates. The repository based on the data elements was mapped to the existing requirements of reports and visualization. An additional data query engine with generic visualization tool was also prepared to ensure ease of use.

4.4.2 It is also important that these data being vital to the functioning of RBI as regulator and supervisor of the financial system, the continued availability and smooth migration to the new system is ensured. Towards this end, the reporting formats were mapped to the element-based repository so that the reporting entities can switch over to the new system in a gradual manner, with the largest and most significant ones at the beginning.

4.4.3 The supervisory and regulatory data templates are quite heavy with thousands of data cells from multiple disparate data elements. Some data templates contain data of more than 20 data elements or DSDs. One of the major challenges were to convert the old database to new SDMX based data series and carrying out data audits. A specific converter application was built to help not only for old data but also to use it for ongoing template-based data which would be reported during transit period to SDMX. Converter tool made as generic engine which will access data template-based database, mapping table and metadata to generate SDMX data series for each reporting entity.

4.4.4 Variety and volume of data collected from about regulated entities can be gauged from the fact that there are about 245 data templates, applicable to different classes of over 10,000 regulated entities. Most of the data templates are in tabular form and each cell in the table corresponds to a particular business data element. Number of cells for a particular data element are based on the multiplication of number of values in column and row dimensions. These templates were designed based on the operational departments' need to view multiple data series together as a part of offsite monitoring. These are the reports which business users use for day-to-day operational activities including inputs for decision making. The same data elements were found in multiple data templates which are collected by different departments for specific purposes. The granularity or dimensionalities were also varied according to its use.

4.4.5 Accordingly, efforts made while designing DSDs to define unique data elements and to reduce reporting burden on regulated entities. Considering all the aspects, over 1,000 DSDs were created along with other artefacts, such as, code lists and attributes. Given the complexity, there is scope for further improvement by reduction in the number of DSDs in the next phase.

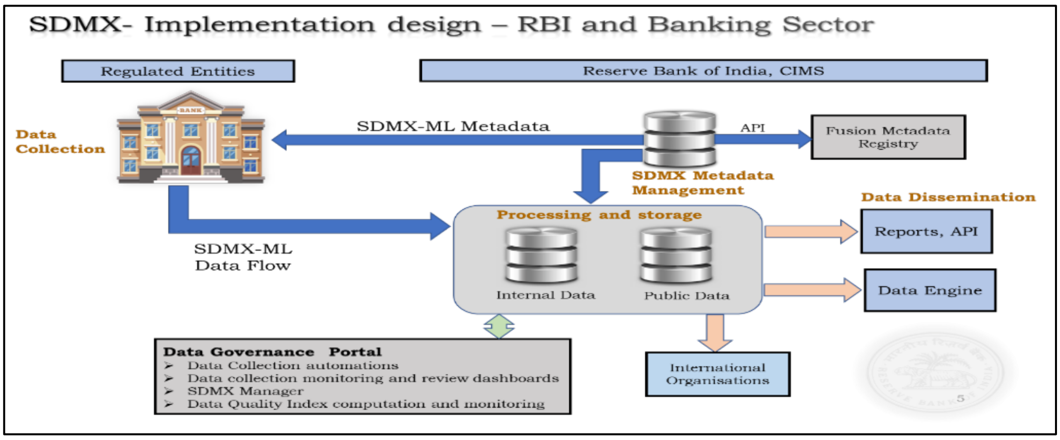
4.4.6 In addition, macro-economic indicators used by the Reserve Bank and disseminated on its public web portal (Database of Indian Economy - DBIE) were also analyzed and another 155 DSDs are prepared. The uniformity of approach for regulatory data as well as other macro-economic and market data ensured that same design and tools will work across the system.

4.4.7 After combining the DSDs with the values of dimensions, for which data are analyzed, an estimated 300,000 data series (without considering entity dimension) are required to be created from these 1155 DSDs. While modelling these combinations from various DSDs, a unique code was allotted to each data series. This unique code is named as Data Model Identification (DMID). For each reporting entity, there will be a separate data series for same applicable DMID, as entity is not part of DMID.

4.5 Implementation architecture:

4.5.1 A high-level conceptual design for creation and management of SDMX artefacts, dissemination of metadata with regulated entities, data collection, data governance and data disseminations, was based on data lake concept using Hadoop based eco-system with Hive database for storage. The application as detailed below was developed, tested, and deployed.

Picture-3: SDMX -Implementation design



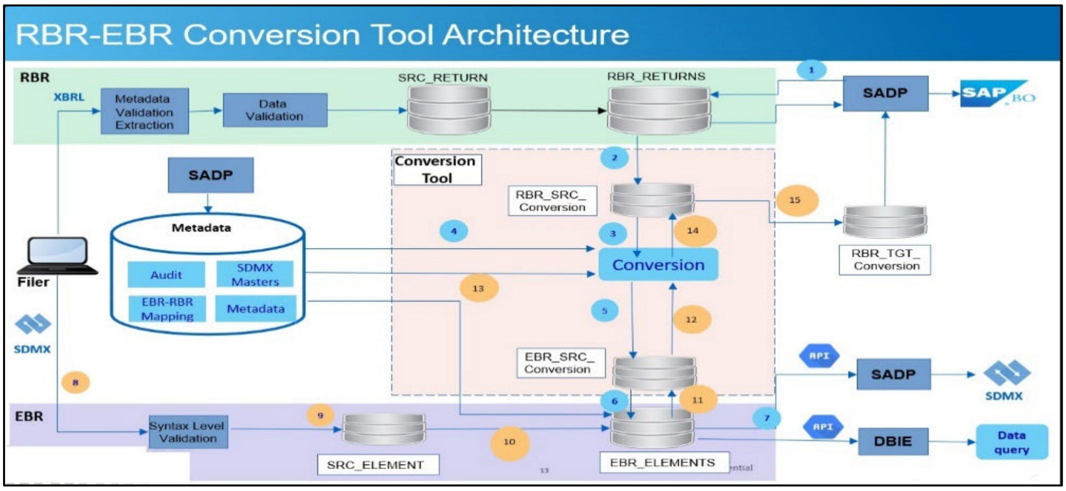
4.6 Metadata Management

4.6.1 A major pain points mentioned by the users was the time consumed in the design, develop, and modify existing data template structure or new data collection templates. Accordingly, a flexible metadata driven data collection and storage architecture was built with technical mapping tables. A specific web-based software module developed to manage and maintain all SDMX artefacts such as DSDs, Code Lists, etc.

4.6.2 The Fusion Metadata Registry (FMR) has been integrated with custom developed metadata management module through Application Programming Interface (API). Technical mapping tables are created with link to metadata artefacts to their storage database table. The system is designed to provide flexibility to make changes to database table structure based on the metadata definitions. This will enable quick implementation of changes in data collection processes.

4.7 Data Collection: The two most important aspects while collecting data are format and data transmission technology. The entities reporting to the RBI are widely different based on their type and maturity level of their Information Technology (IT) Systems. The ADF mechanism mentioned in section III has been used for transmitting SDMX ML data files using server-to-server and API. The entire value chain of granular data integration starts at the reporting entity IT Systems to the ADF server, generating data file from ADF server and transmitting to RBI is automated. As the expertise required to generate SDMX ML files may not be available with all the banks, a custom build converter tool is developed to help convert data to SDMX ML file. While SDMX based data structure and data repository has been implemented in CIMS, the data collection from regulated entities in SDMX format is being taken up in phased manner. Transition to the SDMX based reporting for large banks, which cover over 90 per cent of the banking business is currently underway and is expected to co-exist with the legacy reporting of small and medium sized financial entities over the medium term.

Picture-4: Conversion tool Architecture



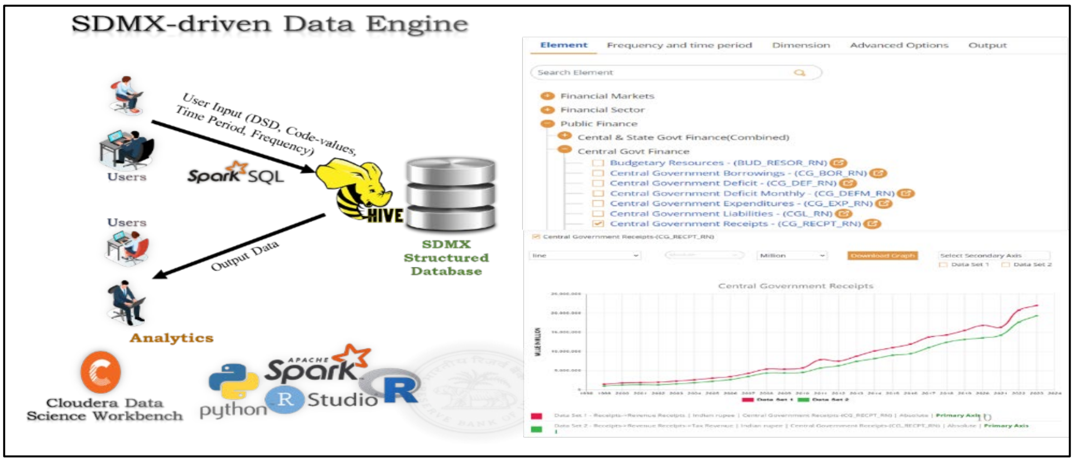
4.8 Data Governance:

4.8.1 The entire data storage structure is based on SDMX standard. Databases are connected to metadata using technical mapping tables. This standardization has helped to implemented data governance related aspects to manage data life cycle, data quality measurement and access management. A specific web portal has been developed for internal users and regulated entities, provides the governance functionality such as data collection channel management, user access management, monitoring data collection process, generating alerts for timely data submission, and dashboards for data submission status.

4.8.2 A common data governance portal also enables measurement of data quality based on important parameters such as timeliness, accuracy, consistency, credibility, relevance, stability, and accessibility. An index (data quality index) to provide a comprehensive measure of data quality has been developed which can be computed for each entity and data element with aggregation at various levels.

4.9 Data Dissemination: Structuring of metadata in SDMX format has enabled the use of metadata in report logic and generalization of report modules. This has enabled development of a generic analytic engine named as Data Query Engine (DQE). This engine is plugged in web portal for providing search on metadata, select DSDs and specific code values as filtering conditions, dimension reduction or aggregation, basic computation on data series and to view data. Data can be viewed as tabular and as a chart. The DQE provides facility to download data in popular formats and charts. The standard metadata driven data storage provide ease of providing data access and permitting advanced statistical analysis.

Picture-5: Data Query Engine



Typical analysis involves use of data from multiple domains, which are required to be selected from different reports and then combined outside the primary data platform. Storage and organization of data in SDMX format in a generic repository enables selection of data from multiple domains, combining data with different periodicity. Selection can be done through the data catalogue or by searching the element. This provides flexibility in analysis and format of output. The time-consuming report generation and manpower overheads for development can be completely eliminated, as long as the generic report formats are acceptable to the users.

V.Implementation Challenges

5.1 Implementation of SDMX repository and multipurpose database design was not without its fair share of challenges. The main objective of achieving consistent data across different domains will be achieved, when the full roll out of the SDMX based reporting is implemented and the existing repository is based on the mapping of legacy data, which were collected with purpose specific formats and definitions and were required to be reconciled across domains.

5.2 The legacy formats were so arranged that many RBI prescribed formats contained several two/three dimensional sheets and each cell in these matrices represented a data cell which had the row and column values as attributes. These nearly 300,000 data cells were required to be analysed to identify the underlying measures and a unified dimension tables, which were required to generated from multiple lists of values with often overlapping coverage. This job required both proper understanding of the data model as well as business requirement which these data were meeting. Manpower with knowledge in both these domains is scarce and was not available to the project in sufficient strength.

5.3 The regulatory institution domain is diverse in size, business focus and technological maturity. Therefore, it was realized early in the project lifecycle that it will not be feasible to implement the new SDMX based reporting across the entire institutional spectrum. The requirement of resources and technical manpower at the reporting entity side is too onerous for mid to small sized institutions. Further, given their small contribution to overall business in value terms, may not justify the deployment such resources. Therefore, it was necessary to design a system, which as a mix of legacy and new reporting formats, where conversion of data from one format to other, in both directions is necessary.

5.4 These challenges were overcome by creating innovative solutions such as splitting the DSDs, introducing dummy dimensions such as “Measure Type” to take care of multiple measures, introducing hierarchy in SDMX manager module, providing parameterized priority for data cells in mapping table to consider single value when there were data mismatches and by building and providing a specific converter tool to the regulated entities.

VI. Concluding observations

6.1 The complexity of information management, particularly when similar data are collected from same cohorts of regulated entities for different purposes and specific definitions had become more prominent, when the data collection and management was centralized over the past decade. Centralization of information system provided an opportunity to examine and overcome these problems comprehensively. Earlier exercises of harmonization of definitions, provided a platform, to develop the SDMX based information ecosystem in the Reserve Bank of India.

6.2 SDMX standard is not only provided mechanism of flexibility and efficiency in collecting metadata and data but also helped to build a simple database architecture for building data repository which can be used for data query, data visualisation and advanced statistical analysis. They also facilitated implementation of data standardization across all the functions, which collect and disseminate data. A data conversion tool developed for generating SDMX time series data from the legacy returns-based data using mapping tables helps in minimising the transition time. This tool has been successfully used in internally in the Reserve Bank and will be rolled out to the regulated entities in the coming months. The generic data element-based reporting by the banks and SDMX based analytical data query and visualisation application will also ensure that the emerging data needs can be met, by merely tweaking the system parameters and dimension values, rather than making any changes in the software modules. This will eliminate the "time to market" to almost zero.

6.3 The CIMS has been in operation since June 2023. Different cohorts of regulated entities such as scheduled commercial banks (SCBs), urban co-operative banks (UCBs) and non-banking financial companies (NBFCs) were progressively onboarded for reporting on the new CIMS system. Reserve Bank has conducted numerous training sessions for officials of regulated entities, complemented by one-to-one handholding sessions to ensure that they are able to report the data in the new CIMS system[5], without facing much difficulty.

----

[5]https://rbidocs.rbi.org.in/rdocs/Speeches/PDFs/STATISTICSDAYCONFERENCE722C82A17B374C37A733A3BD8BA1633C.PDF

References

[1] Das, S (2023) - Inaugural Address at the 17th Statistics Day Conference organised by the Department of Statistics and Information Management, Reserve Bank of India, Mumbai, June

[2] Das, S (2024) - Inaugural Address at the 18th Statistics Day Conference organised by the Department of Statistics and Information Management, Reserve Bank of India, Mumbai, June

[3] George, L., 2011. HBase: the definitive guide. O'Reilly Media Inc.

[4] List of Returns submitted to RBI – www.rbi.org.in (Regulatory Reporting)

[5] Olston, C., Reed, B., Srivastava, U., Kumar, R., Tomkins, A., 2008. Pig Latin: A Not-So-Foreign Language for Data Processing. In 2008 ACM SIGMOD international conf. on Management of data.pp.1099-1110

[6] Patra, M. D. (2024): Keynote address at the SAARCFINANCE seminar on Emerging Technologies in Central Banking and Finance, Goa, January

[7] Reserve Bank of India (2010): Approach Paper on Automated Data Flow from Banks.

[8] Reserve Bank of India- Annual Report 2023-24.

[9] SDMX standard (https://sdmx.org/)

[10] Shvachko, K., Kuang, H., Radia, S., Chansler, R., 2010. The Hadoop distributed file system. In MSST'2010, IEEE 26th Symposium on Mass Storage Systems and Technologies, pp. 1-10

[11] White, T., 2009. Hadoop: The Definitive Guide, first ed., O'Reilly Media, Inc.

[12] Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I., 2010. Spark: Cluster Computing with Working Sets. In HotCloud'10, 2nd USENIX conference on Hot Topics in Cloud Computing (pp. 1-7).

Implementing Statistical Data Metadata Exchange (SDMX) in Centralised Information Management System – Challenges, Solution Design and Benefits

Debasis Nandi, and Prakash Madda

12th IFC Biennial Conference, BIS Basel,
22-23 August 2024

The views expressed in this paper are of the authors and not the institution to which they belong

Information management- Reserve Bank of India

RBI collect, compile, and disseminate wide range of data

245 data templates, applicable to different classes of over 10,000 regulated entities

Monetary policy surveys, Government agencies and internal systems

Diversity in data characteristics and their diverse usage across departments

Steps taken -harmonise the concepts and definitions and rationalise the reporting system

Surge in data volume, variety, and velocity in new modern world

Technological obsolescence and scalability limitations

Comprehensive strategy necessitated :

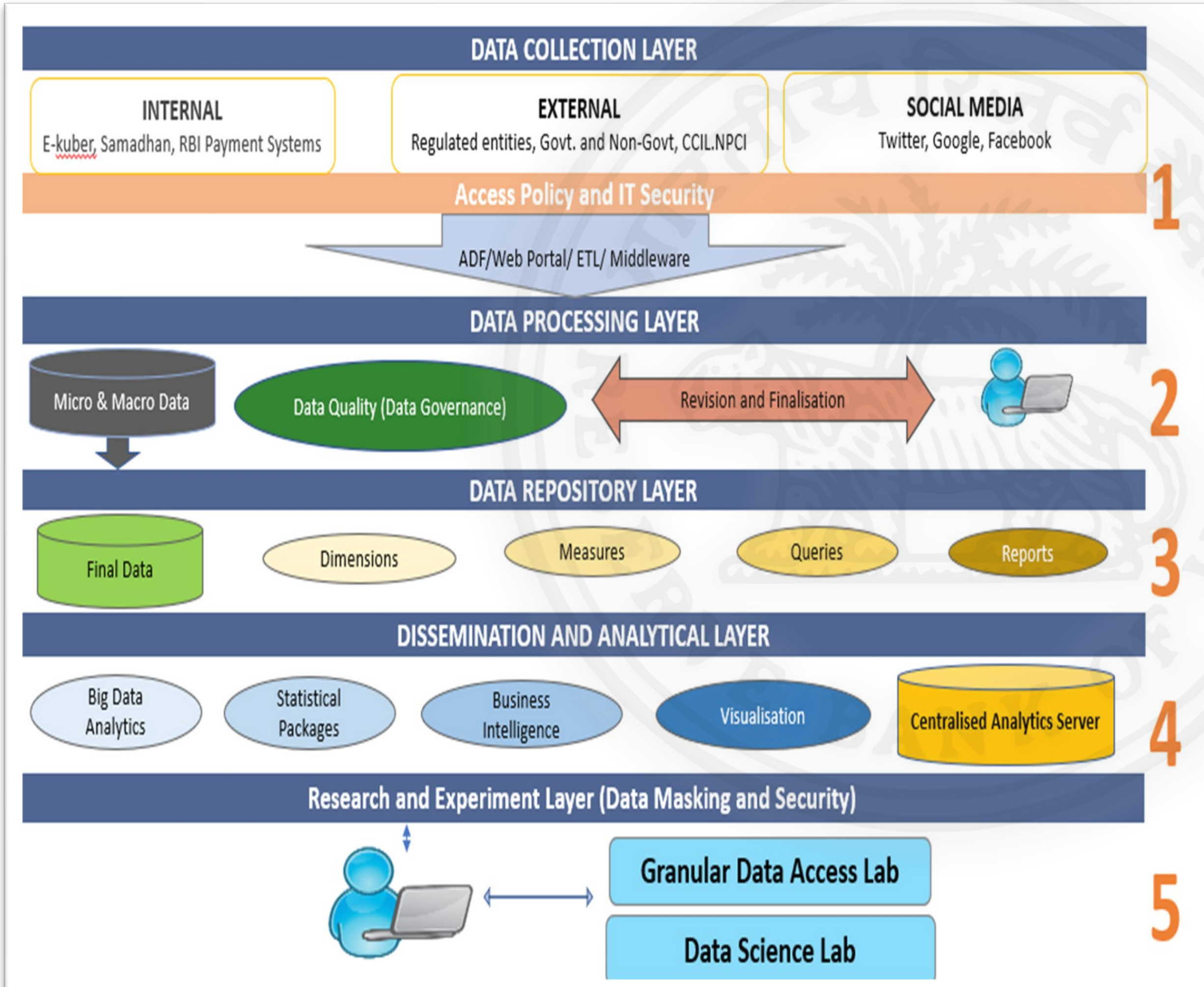
a. To overhaul entire value chain of data, from collection, modelling, storage, retrieval and usage by the end-user.

b. Implement Data Lake (big-data Hadoop technology).

c. Adopt international standard facilitating the exchange of statistical data and metadata.

Centralised Information Management System (CIMS)– The Big Shift

Conceptual Architecture



CIMS major features are:

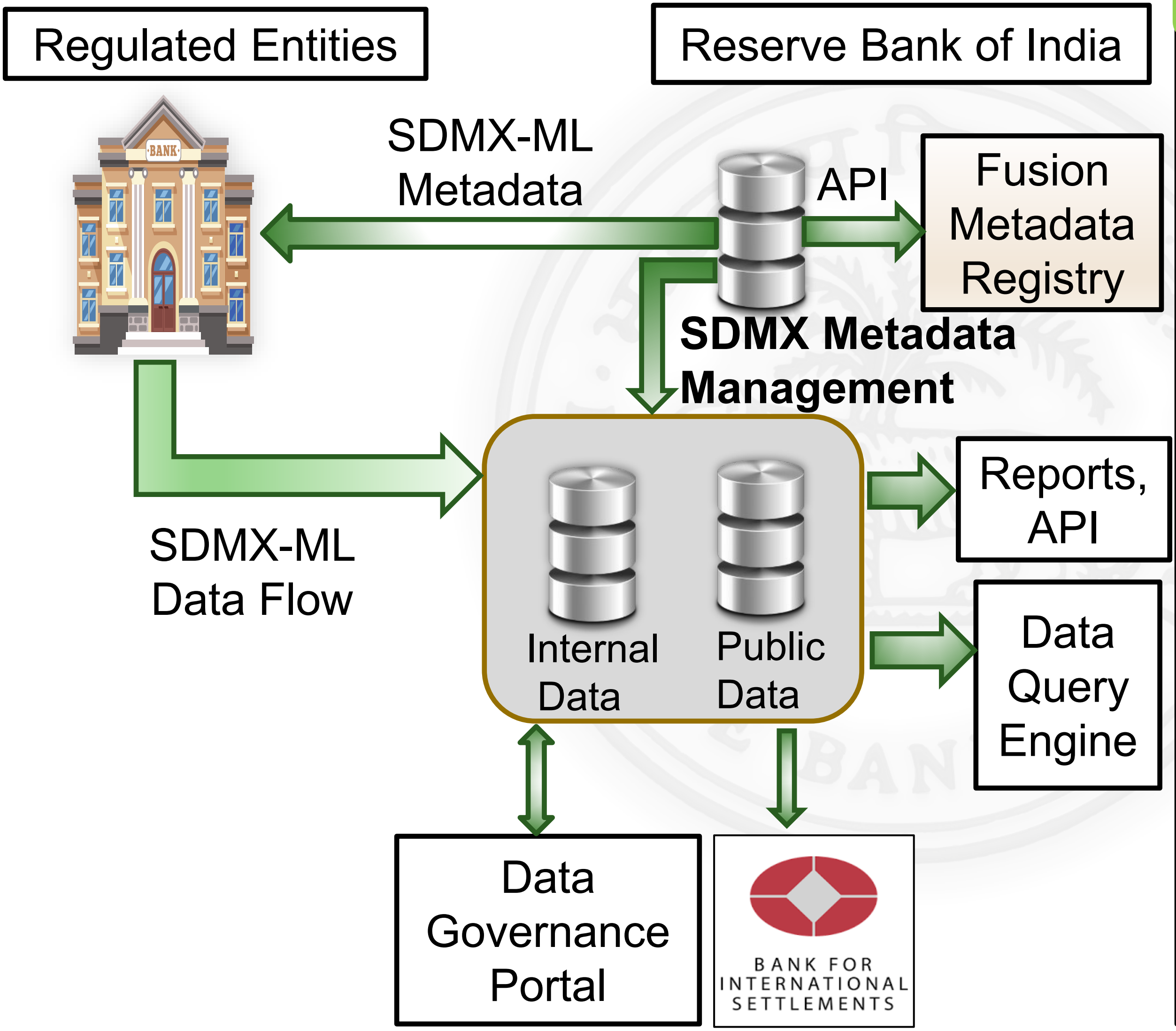
a. Automation of data transmission

b. Development of state-of-the-art SDMX data architecture;

c. Management of high volume structured and unstructured data using Bigdata open-source technology like Hadoop;

d. Development of integrated data analytics, visualization, and data dissemination

SDMX – Implementation Design



Salient features are:

- **Flexible metadata driven data collection and storage architecture was built with technical mapping tables**
- **300,000 data series estimated from 1155 DSDs.**
- **Unique code Data Model Identification (DMID) was allotted to each data series.**
- **Web-based software module developed to manage and maintain all SDMX artefacts such as DSDs, Code Lists etc.**

SDMX – Data Mapping (Template → SDMX)

TOTAL CELLS : 2938

MAPPED CELLS : 2938

EBR VERSION: 1.0.0

Valid From Date: 01/01/2021

Description: Publish

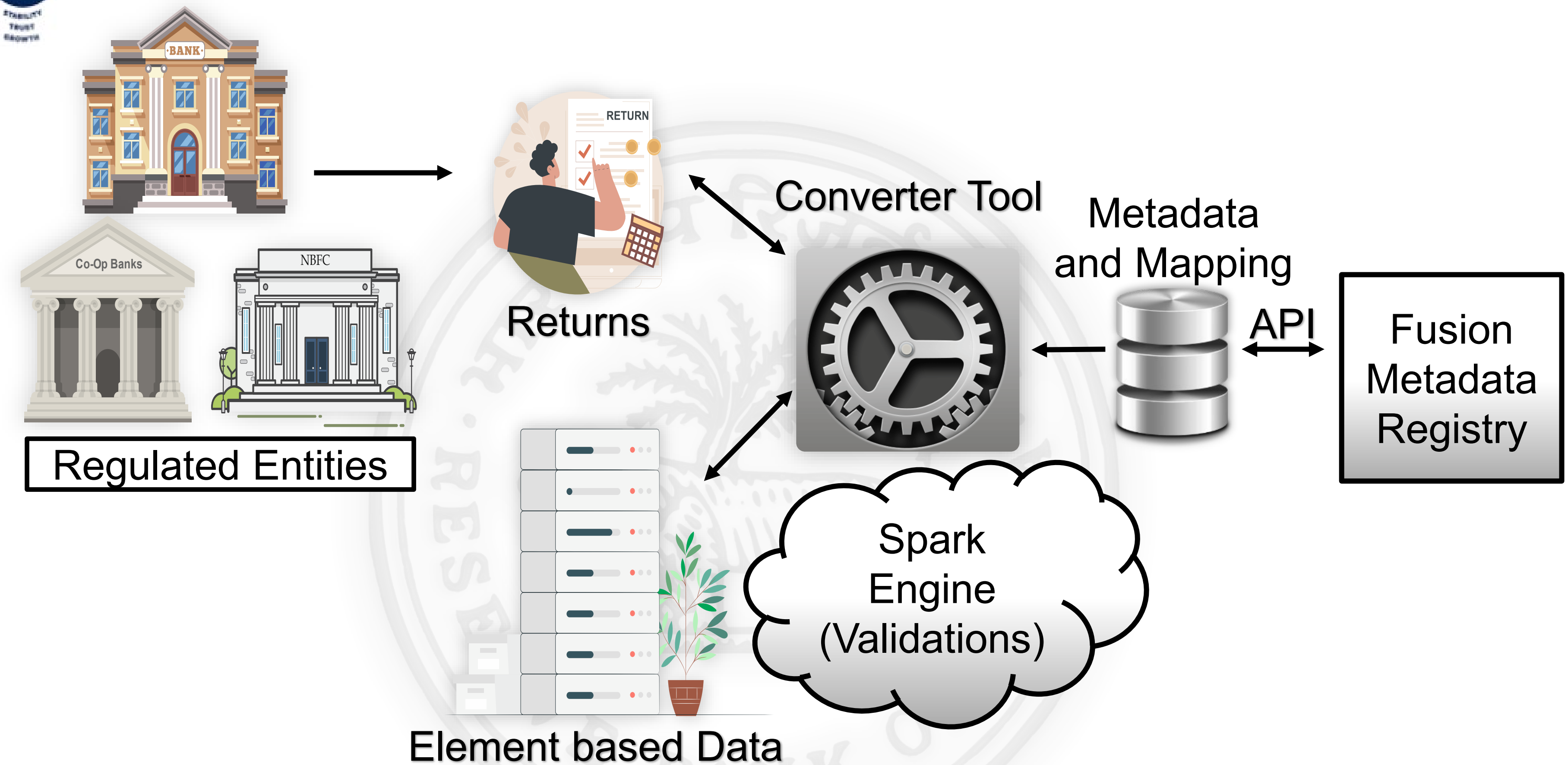
Deactivate

Sec1Part-A | Sec1Part-B | Sec1Part-C | Sec2Part-A | Sec2Part-B And C | Sec2Part-D1E1F1 | Sec2Part-D2E2F2 | Sec2Part-G1A

Section 1: Assets and Liabilities						
Part-A: ASSETS (Amount Outstanding at end of Month)	Domestic Operations					
	Total					
I.Cash Funds	1					
I.1 Cash on Hand	7	8		10	11	12
I.2 Balances/Deposits with RBI/central Banks	13	14		16	17	18
II.Due from Banks in India	19	20	21	22	23	24
II.1 Balances in Current a/c's (Annexure 1- A)	25	26	27	28	29	30
II.2 Money at call and short notice	31	32	33	34	35	36
II.3 Placements/ Time Deposits	37	38	39	40	41	42
II.4 Loans and advances	43	44	45	46	47	48
III. Due from FIs/Central Counter Parties (CCPs) in India	49	50	51	52	53	54
III. 1 Due from FIs in India	55	56	57	58	59	60
III. 2 Due from CCPs in India (incl. CBLO Lending)	61	62	63	64	65	66
III. 3 CBLO Lending Repayable within 15 days	67	68	69	70	71	72
III. 4 CBLO Lending Repayable beyond 15 days	73	74	75	76	77	78
IV. Due from Overseas Banks	79	80	81	82	83	84
IV.1 Balances in current accounts (Annexure 1-B)	85	86	87	88	89	90

[ELEMENT];
LNA(3.0)
#
[CODELIST DIMENSION];
LN_LI | DUEFRMBANKS_LNA;
AREA_OPERATION | DOM_OP;
MEASURE_TYPE | TOTAL;
AQC | N_A;
CURRENCY | INR

SDMX – Novel Conversion tool

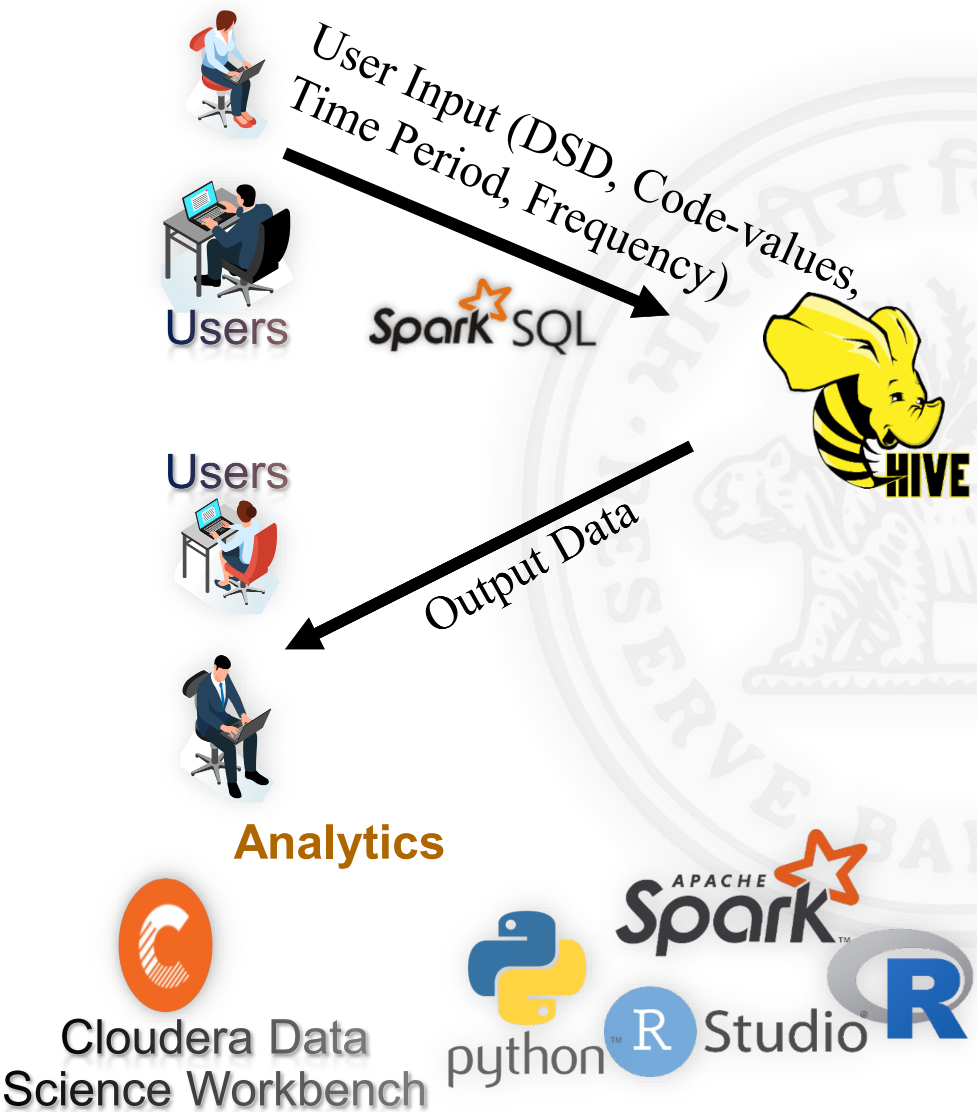


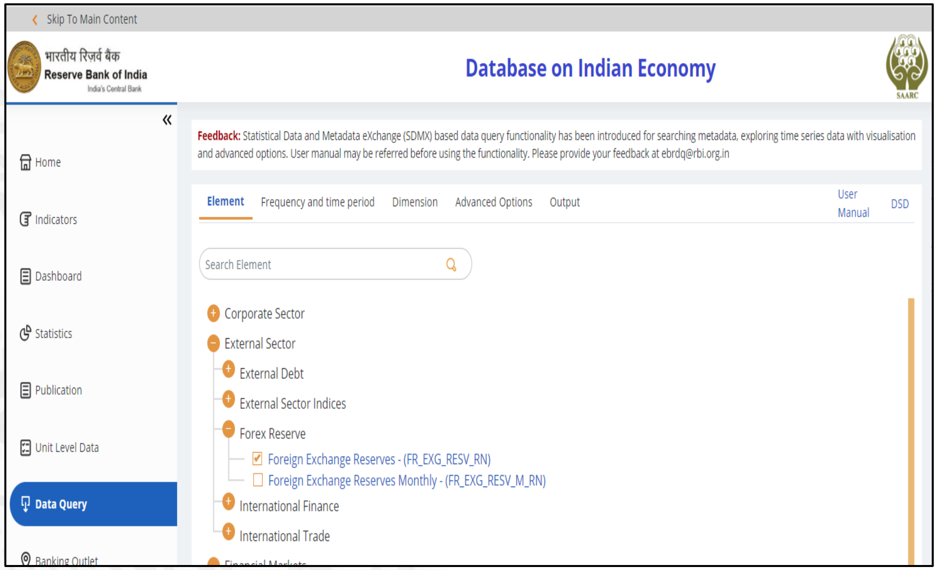
Converter tool made as generic engine

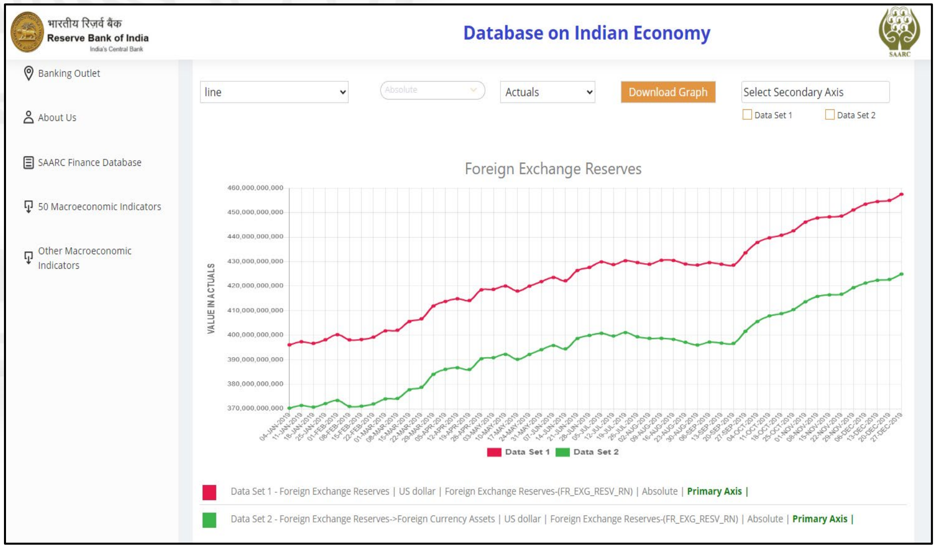
1,000 Elements and 1,200 Dimensions with Code Lists from 245 Returns

155 Elements and 217 Dimensions from Macro-Economic

SDMX – Data Query engine







Implementation Challenges

300,000 data cells were required to be analysed to identify the underlying measures and a unified dimension tables

Manpower with knowledge in proper understanding of the data model as well as business requirement is scarce and in sufficient strength available to the project

Regulatory institution domain is diverse in size, business focus and technological maturity.

Necessary to design a tool for conversion of data from one format to other, in both directions.

Challenges were overcome by creating innovative solutions such as splitting the DSDs, dummy dimensions , hierarchy in SDMX manager module, converter tool etc.

Concluding observations

A data conversion tool developed for generating SDMX time series data using mapping tables helps in minimising the transition time.

SDMX standard helped to build a simple database architecture for building data repository which can be used for data query, data visualisation and advanced statistical analysis.

CIMS has been in operation since June 2023. Different cohorts of regulated entities such as scheduled commercial banks (SCBs), urban co-operative banks (UCBs) and non-banking financial companies (NBFCs) were progressively onboarded for reporting on the new CIMS system.

Reserve Bank has conducted numerous training sessions for officials of regulated entities, complemented by one-to-one handholding sessions to ensure that they are able to report the data in the new CIMS system, without facing much difficulty

Thank You