Paper SAS3123-2019

No Time to Create an Enterprise Data Warehouse? No Problem. Blockchain to the Rescue!

Angela Hall, SAS Institute Inc.

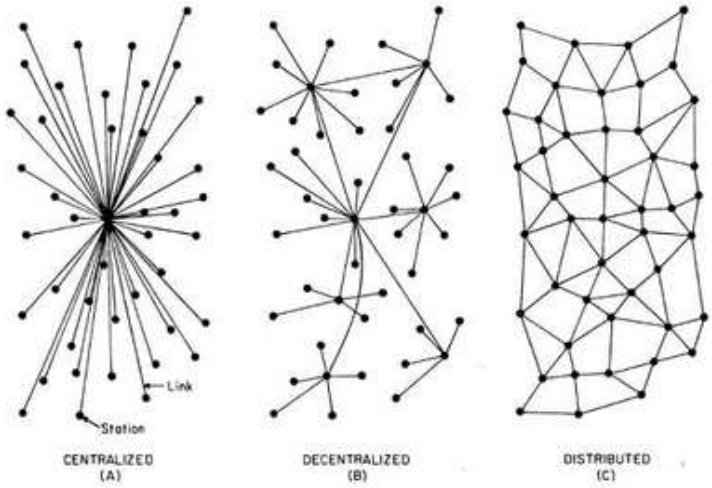
ABSTRACT

A majority of analytic work cannot start until an enterprise data warehouse (EDW) exists or significant database development work occurs. This time-consuming and usually cost-prohibitive prerequisite gets in an organization's way of generating real analytical value quickly from data. Other issues with EDWs include security (single-point-of-failure), scalability, data transfer and resource constraints, and accessibility. What if blockchain technology could help an organization skip the EDW process altogether? And how can SAS® be leveraged to generate immediate value in ensuring valid blockchain creation as well as timely analytic results? In this paper, we explore several potential cross-industry use cases for blockchain and how SAS can assist in creating a block and generate value from the block.

INTRODUCTION

Probably everyone has now heard of bitcoin. If not, it is "a type of digital currency in which a record of transactions is maintained and new units of currency are generated by the computational solution of mathematical problems, and which operates independently of a central bank" (Google 2019). Most consider blockchain and bitcoin to be synonymous. However, blockchain is simply a mechanism of data storage and the backbone for bitcoin.

The figure below (from a 1964 paper on communication networks) describes three different network systems: centralized, decentralized, and distributed. This is commonly used to describe blockchain networks as well. Before we go there, let's relate this to data warehouses and data storage ideas. First, a centralized system is an enterprise data warehouse (EDW) where everything is merged and stored in a single-point-of-truth (and single-point-of-failure) environment. In a decentralized system, multiple entities have copies of the data but can link out to other (and different) data stores. Finally, a distributed system has no data stores; all the nodes contain their own data but can join across systems.



Centralized: Enterprise Data Warehouse
Decentralized: Multiple Copies of Data with links out to different data stores
Distributed: Various data elements are stored in different systems

Figure 1. Paul Baran's Systems Diagram and Proposed Relationship to Data Storage

GAINING VALUE USING BLOCKCHAIN

When we consider our client's consulting services projects, the majority of analytic work is gated by the need to first create EDWs. These time-consuming and usually cost-prohibitive projects get in the client's way of generating real analytical value quickly from data. Other issues with EDWs include security (single-point-of-failure), scalability, data transfer and resource constraints, and accessibility. What if blockchain technology could help a client **skip** the EDW process altogether and allow SAS® to generate value in ensuring valid blockchain creation as well as timely analytic results?

WHAT IS BLOCKCHAIN?

Before I suggest that blockchains will alleviate the cost and risk of establishing an EDW, let's take a moment to describe blockchains. Most individuals know of bitcoin; it is by far the most famous use case of blockchain technology. How it works is that each coin is a new block on the blockchain and data is added to the block as the coin is transferred between owners. The data about this block is replicated in a ledger that is stored throughout multiple nodes in a decentralized network. The blocks are unalterable due to the exorbitant cost (compute power) required to hack into each node's copy of the ledger.

Consider first that not all data needs to be stored in a blockchain. For example, placing all health-care claims data into a blockchain (each claim being a new block) to document each change to that claim appears to be a replication of the claims transactional database. However, tracking a member's location (address) from various sources (claims and eligibility) could be useful for various reporting and fraud analytic scenarios.

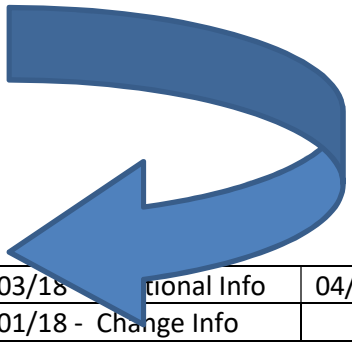
SCD VERSUS BLOCKCHAIN

In a Type-2 slowly changing dimension (SCD), data is appended to a table and if new data appears for a specified ID then an end-date is added to old records when updated data records are appended. Tracking an ID and the changes and updates requires pulling records by that ID, and then transposing is required to complete calculations such as events over time. However, in a block, information is stored by a specified ID and data is appended to

the block as additional characters and text on that specified ID. This wide table design allows for faster analytics when running scenarios that require reviewing events for each specified unique ID.

SCD Table

ID	Date	Event
01023	02/21/18	New Record
32488	03/11/18	Additional Info
32488	04/01/18	Change Info
01023	04/03/18	Additional Info
01023	04/10/18	Void Info



Blockchain

Block: 01023	02/21/18 - New Record	04/03/18 [illegible]tional Info	04/10/18 - Void Info
Block: 32488	03/11/18 - Additional Info	04/01/18 - Change Info	

Figure 2. SCD Versus Blockchain Layout

HOW CAN SAS BE INVOLVED?

SAS® Event Stream Processing provides the ability to complete analytical models of data in-flight. In the high-level process flow below, SAS Event Stream Processing receives the incoming data stream and completes various routines such as data quality, entity resolution steps, and so on. It then updates an existing block or creates a new one.

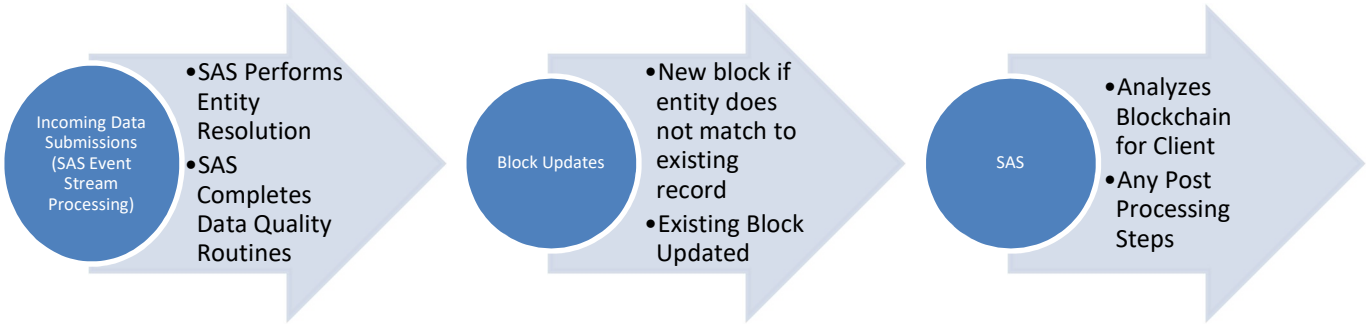


Figure 3. Process Flow of Block Chain Creation

SAS EVENT STREAM PROCESSING

As an overview, SAS Event Stream Processing receives incoming streams and can execute processes and analytics to improve analytic models as well as complete data enrichment steps. The architecture of this product allows for real-time continuous monitoring, eliminating data processing maintenance windows.

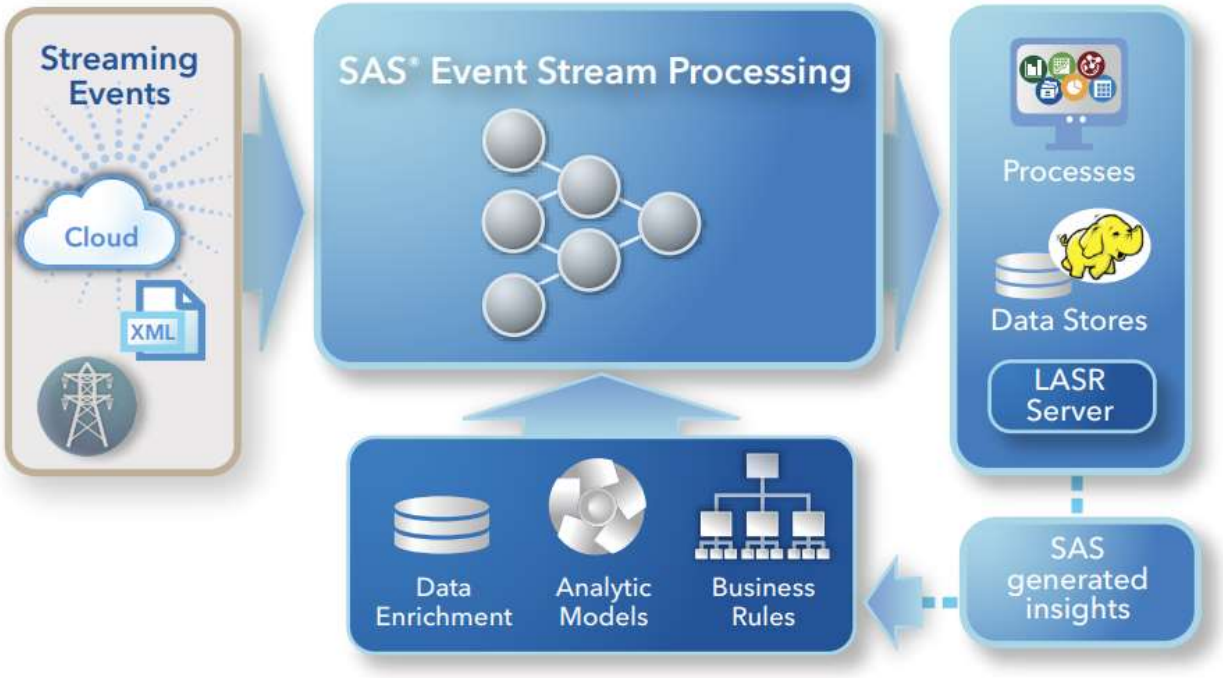


Figure 4. SAS Event Stream Processing Model

USE CASE 1 CRIMINAL JUSTICE

A lot of processing--and therefore data delays--can occur during the centralized EDW data loading process to combine records. The cost of maintaining the EDW is large but it is priceless for law enforcement to have access to up-to-the-minute information on criminals whose information they are directly in contact with.

Using blockchain could speed the processing of data from overnight to minutes by having a block for each criminal. Each participating data source (court systems, county jails, DMV, fish and wildlife, and so on) would modify the process for data submission from nightly/weekly/monthly frequency to an API submission for each new record.

SAS would receive the incoming data submission, and immediately determine whether the individual is an existing block in the system to which it would add the information or necessitate the creation new block.

SAS would then be used to analyze the blocks and provide intelligence and reporting back to law enforcement officials.

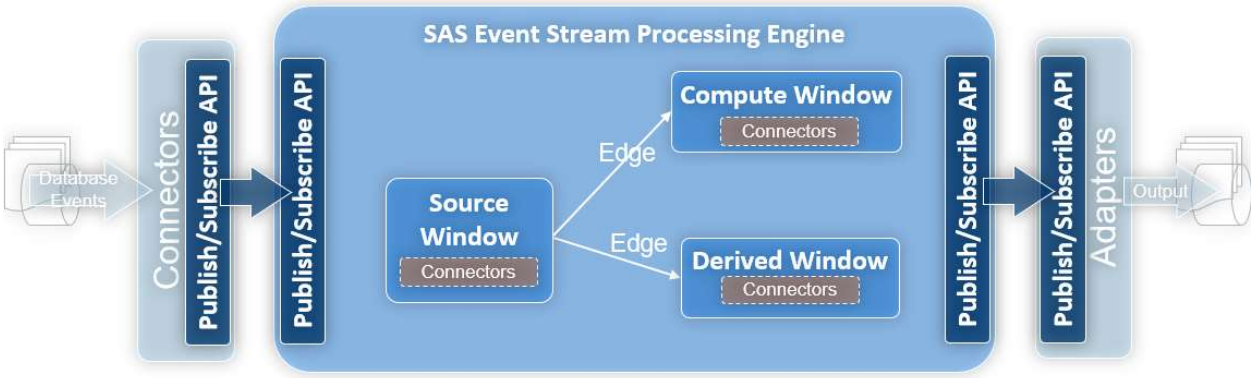


Figure 5. SAS Event Stream Processing Engine Example

USE CASE 2 HEALTH-CARE ELIGIBILITY

One of the common issues with health-care analytics is understanding an individual's eligibility for Medicare, Medicaid, and so on, over time as the member gains access to different levels of care. If each member were a block in the blockchain, data could be added to the block as the member changes status in the various systems. Again, SAS would be useful in cleaning data and completing entity resolution on the member prior to adding or creating a block.

SAS would then be used to analyze the block and enable health-care offices to better understand its member community. In the below example, blocks are organized by the member, and calculations between events can occur easily for each individual.

Block: John Doe	1/2018 – Eligible	2/2018 – Hospital Care	2/2018 – Prescription
Block: Jane Doe	2/2019 – Ineligible	4/2019 – Eligible	5/2019 – Childbirth

Figure 6. Block Formatted Member Eligibility Data

USE CASE 3 TRANSPORTATION TRACKING

State transportation agencies are attempting to track events and assets on each identified road segment. In an EDW, the data is tabular and would need to be transposed to a segment prior to analytics. However, if the data for each asset and event is added to a road segment's block in a blockchain, the processing time is significantly reduced when not using the "by segment" grouping of data.

CONCLUSION

While the hype of bitcoin has lessened, the benefit of using blockchain as a data storage mechanism has its merits. As discussed, the immediate processing and tracking of events can significantly reduce the implementation time seen in traditional data warehousing projects and can generate true analytic value.

REFERENCES

Baran, Paul. 1964. "On Distributed Communications: I. Introduction to Distributed Communications Networks." Rand.org. Available https://www.rand.org/content/dam/rand/pubs/research_memoranda/2006/RM3420.pdf. Accessed August 27, 2018.

"Bitcoin." Google Dictionary. Available https://google.com. s.v. "dictionary." Accessed February 26, 2019.

"Blockchain." Wikipedia. Available https://en.wikipedia.org/wiki/Blockchain. Accessed August 27, 2018.

Penfield, Sam. "A Practical Approach to Blockchain Analytics." Available https://blogs.sas.com/content/sascom/2017/12/15/practical-approach-blockchain-analytics/. Last modified December 15, 2017. Accessed August 27, 2018.

Tar, Andrew. "Decentralized and Distributed Databases, Explained." Cointelegraph.com. December 2, 2017. Available https://cointelegraph.com/explained/decentralized-and-distributed-databases-explained. Accessed August 27, 2018.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Angela M. Hall
Director, Consulting
SAS Global Hosting and U.S. Professional Services
Angela.Hall@sas.com