ProvenDB

Meeting APRA Data Risk Guidelines with Google BigQuery and ProvenDB

Meeting APRA Data Risk Guidelines with Google BigQuery and ProvenDB

Background

The APRA (Australian Prudential Regulation Authority) Data Risk guidelines clearly articulate a high bar for data integrity (our emphasis):

> *Auditability (the ability to* ***confirm the origin of data*** *and provide* ***transparency of all alterations****) is a key element to verifying data quality. It involves the examination of data and associated audit trail, data architecture and other supporting material. APRA envisages that a regulated entity would ensure that* ***data is sufficiently auditable*** *in order to satisfy the entity's business requirements (including regulatory and legal), facilitate independent audit, assist in* ***dispute resolution (including non-repudiation)*** *and* ***assist in the provision of forensic evidence*** *if required*[1]

Most regulated entities struggle to meet these APRA guidelines, primarily because commonly used data storage solutions are unable to provide immutability, full auditing or strong guarantees of data integrity. When attempting to prove or determine regulatory compliance, there might be no indisputable resolution mechanism for disputes relating to data accuracy or to definitely prove the absence of tampering or other faults of commission or omission.

Blockchain technologies represent a quantum leap forward for data ownership, transparency, and tamper resistance. Data written to a reliable public blockchain meets all of the APRA Data Risk guidelines. Unfortunately, such public blockchains cannot provide the necessary functionality, performance or economics required for the storage of large-scale datasets. Public blockchains have highly limited storage capabilities and very poor economics. For these reasons, entities wishing to fully meet APRA guidelines are motivated to combine blockchain capabilities with traditional database technology.

ProvenDB technology

ProvenDB combines the functionality, performance and economics of traditional database services with the immutability and proof of public blockchain solutions.

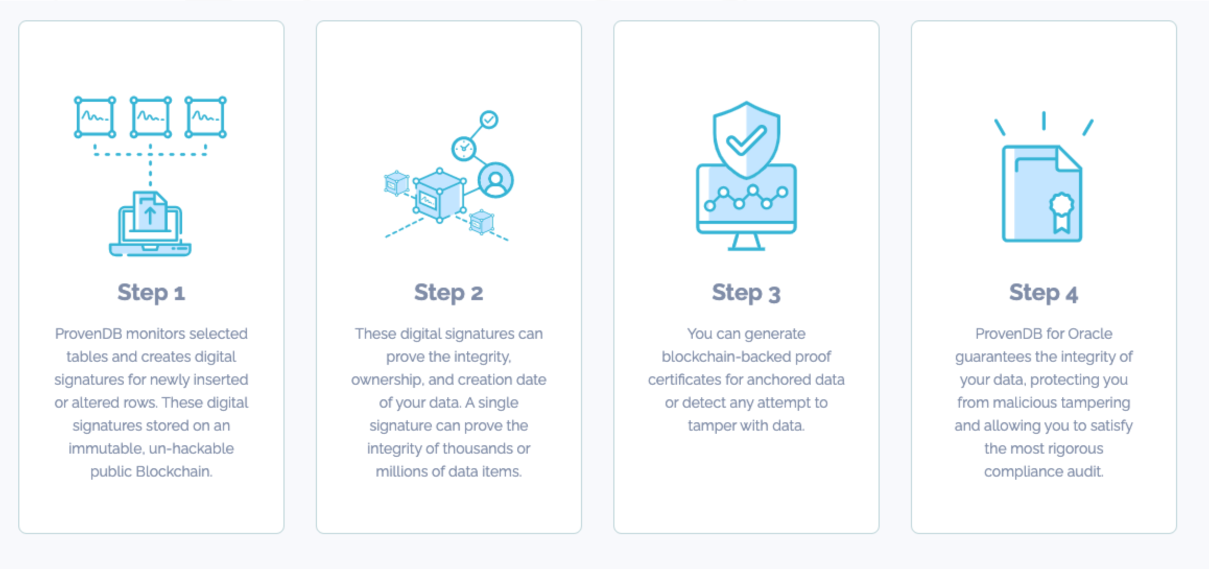
ProvenDB offers the following solutions:

[1] https://www.apra.gov.au/sites/default/files/Prudential-Practice-Guide-CPG-235-Managing-Data-Risk_1.pdf

- The ProvenDB database service is an on-premise or cloud-based MongoDB-compatible database service which provides complete traceability and immutable data storage. Data stored in ProvenDB is anchored to a public or private blockchain using industry-standard cryptographic proofs providing complete evidence of origin and provenance.
- ProvenDB Compliance Vault is a document management system based on ProvenDB, which allows for the easy storage of compliance-related documents. Documents can be stored in the Compliance Vault via the web, email or REST interfaces.
- ProvenDB database adaptors allow data in existing databases – such as Oracle or Postgres – to be anchored to a public or private Blockchain, providing proof of integrity and provenance of existing data.

How it works

When data is added to or modified in a database being monitored by ProvenDB, cryptographic signatures of the data are created. These signatures can be "signed" by your company's cryptographic key (possibly the same key that guarantees the identity of your website). These signatures are aggregated and anchored to a public Blockchain such as Bitcoin, Hedera or Ethereum, or to a private Blockchain such as HyperLedger.



Step 1

ProvenDB monitors selected tables and creates digital signatures for newly inserted or altered rows. These digital signatures stored on an immutable, un-hackable public Blockchain.

Step 2

These digital signatures can prove the integrity, ownership, and creation date of your data. A single signature can prove the integrity of thousands or millions of data items.

Step 3

You can generate blockchain-backed proof certificates for anchored data or detect any attempt to tamper with data.

Step 4

ProvenDB for Oracle guarantees the integrity of your data, protecting you from malicious tampering and allowing you to satisfy the most rigorous compliance audit.

Figure 1 ProvenDB architecture

Once anchored to the public Blockchain, the signatures form an impeccable and irrefutable proof of the integrity and origin time of the data concerned. The Blockchain record – which cannot be altered by any known technology – proves the overall integrity and timestamp of items in the database, eliminating any possibility of undetected tampering or backdating.

Using ProvenDB with BigQuery

Google BigQuery is a cloud-based data warehousing solution suitable for very large datasets which supports SQL language queries and Machine Learning capabilities.

BigQuery itself does not provide any of the native capabilities that would be required to satisfy the APRA guidelines. Data added to BigQuery can be updated at any time, and after a seven-day period, any record of the previous state of the data will be removed. BigQuery does not timestamp data as it is inserted, which means that it is not possible to be certain that a data item has not been overwritten since its original ingestion.

However, by combining the capabilities of ProvenDB and BigQuery, we can create a data storage solution that provides blockchain-backed guarantees of data provenance. Figure 2 shows a possible architecture for such a solution.

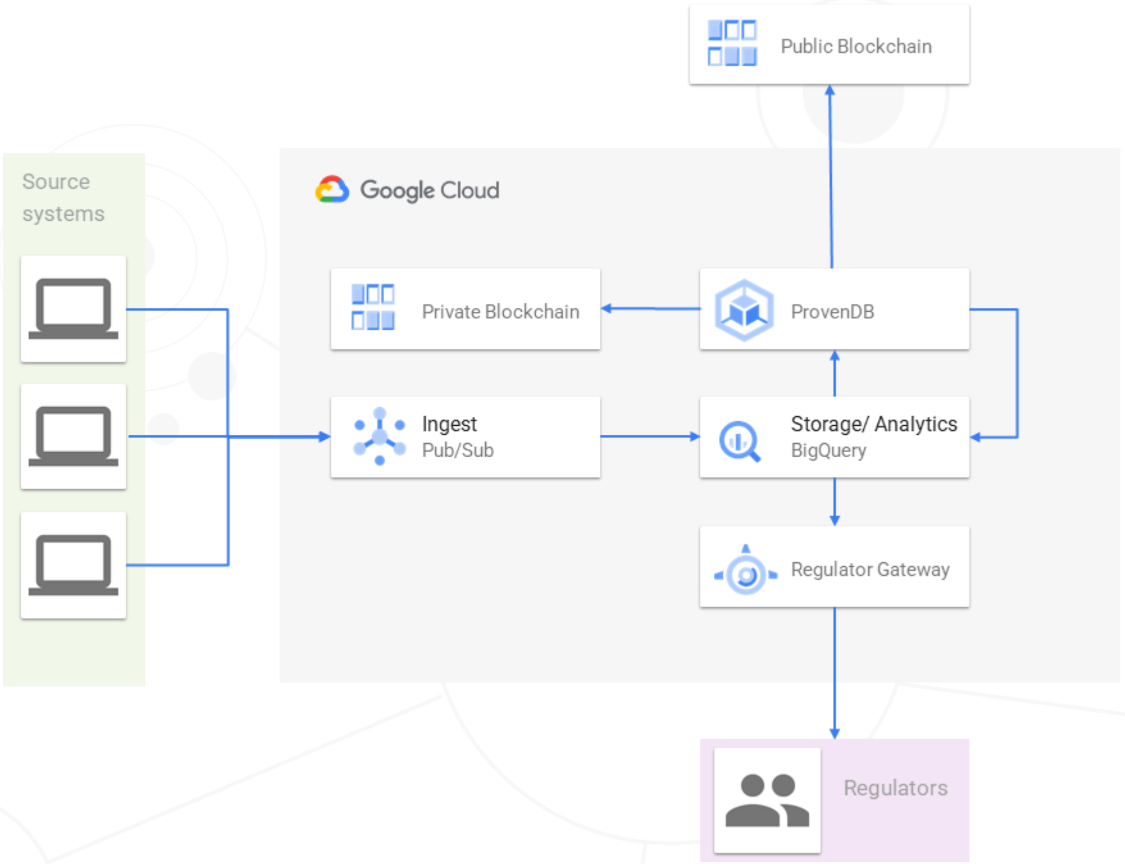


Figure 2 ProvenDB with BigQuery

In this example, multiple systems of records forward data to the BigQuery datastore using Googe Pub/Sub. In some implementations, this ingestion engine might perform data cleansing and other data processing.

It is desirable, though not essential, for the BigQuery schema to be constructed to allow for the efficient scanning of temporal data, for instance, by using a timestamp-based or ascending primary key or by using time-based partitioning in the target table.

ProvenDB periodically scans the target tables using the AS OF SYSTEM time construct to obtain a snapshot of target table data. ProvenDB will also use the built-in SHA256 hash function to calculate a hash value for each row. At this point, ProvenDB will ensure that no primary key has had a hash change since the previous snapshot, e.g., that no data has been incorrectly modified. All modifications in this database must be stored as new "versions" of an existing row so as to preserve the transparency of alterations as required by APRA. If a row has been improperly altered, the previous version of data can be retrieved by the AS OF SYSTEM time construct and restored.

For all new data, ProvenDB will construct a Merkle tree based on the hash values of the new elements and store the root for that Merkle tree on a public or private Blockchain. The resulting cryptographic proof – consisting of the Merkle tree together with blockchain transaction details, will be stored into BigQuery as a proof certificate.

For any row in the BigQuery database, an interested party would be able to calculate the hash of the row and find the matching hash in the ProvenDB proof certificate. Using standard cryptography, the hash could be proved to have been included in the blockchain transaction, and therefore its provenance established.

Private and Public Blockchains

Enterprises with a strong emphasis on cybersecurity often have policies that restrict data transmissions to the public internet. These organizations are often attracted to private Blockchain-based solutions since they do not violate these security policies.

ProvenDB supports private Blockchains as easily as public blockchains. However, there are significant advantages to a public Blockchain-based solution:

- The "root hash" stored on the Blockchain contains no meaningful data. There is no possible security concern related to this data item.
- Private blockchains do not provide the high level of trust created by a public blockchain. Entires in a private blockchain could be falsified by collusion by a relatively small number of parties – sometimes by only one party. In contrast, there is no known case of falsification of data ever on the Bitcoin or Ethereum blockchains.
- Public blockchains such as Bitcoin and Ethereum have proved to be stable and highly available for over a decade. The chance that human or other error might cause the failure of a private blockchain is far more likely.
- The longevity of a proof on a public blockchain that supports a large economic value is relatively assured. However, if a private blockchain is decommissioned, the proofs may be nullified.

For all these reasons, we recommend the use of a public blockchain such as Bitcoin, Ethereum or Hedera. However, from a technical point of view, we fully support private blockchains such as HyperLedger or Quorum.

Conclusion

The most common issue for regulated entities is not evidence of non-compliance; rather, it's an inability to *prove* compliance. For instance, in the Financial Services Royal Commission, the Commissioner noted that certain big banks could not "readily identify how, or to what extent, the entity as a whole was failing to comply with the law." The consequences of the Royal Commission are, of course, well known. It's widely expected that upcoming investigations will find similar issues across almost all industries.

Maintaining compliance documentation in a central data store is an obvious measure that an organization can establish. In this paper, we've seen how such a central store could be established using Google BigQuery and ProvenDB.

Eliminate the costs and risks involved in regulatory compliance with ProvenDB.
Visit www.provendb.com to sign up for a ProvenDB cloud service, or email us at support@provendb.com
to explore options for deploying ProvenDB Compliance Vault within your organization.

ProvenDB

support@provendb.com
www.provendb.com

f 🐦 in M /provendb