Afro-Asian Journal of Scientific Research (AAJSR)
المجلة الأفروآسيوية للبحث العلمي
E-ISSN: 2959-6505
Volume 4, Issue 2, 2026
Page No: 296-305
Website: https://aajsr.com/index.php/aajsr/index
SJIFactor 2024: 5.028 ISI 2025: 0.915 معامل التأثير العربي (AIF) 2025: 0.76

Random Sample Integrity Tests in big, Segmented Databases

Salem Abdulali Ahmed Abdulali*
General Department, Faculty of Economics / Misrata University, Misrata, Libya

إطار عملي لاختبار سلامة العينة العشوائية في قواعد البيانات الضخمة المجزأة دراسة تجريبية باستخدام المحاكاة الإحصائية

سالم عبد العالي أحمد عبد العالي*
القسم العام، كلية الاقتصاد، جامعة مصراتة، مصراتة، ليبيا

*Corresponding author: s.abdulali@eps.misuratau.edu.ly

Received: March 17, 2026	Accepted: May 02, 2026	Published: May 17, 2026

Abstract:
With the rapid expansion of data volumes and the increasing reliance on distributed databases, the integrity of random sampling in statistical analyses has become a significant methodological challenge. Data skew, or uneven distribution, can lead to statistical bias affecting the accuracy of estimates and the reliability of analytical results. This study aims to analyze the impact of this skew on sample representation in distributed environments and propose a practical framework for testing sampling integrity in large, fragmented databases. The study proposes a methodological framework known as the Reliable Sampling Framework (RSF), which combines stratified sampling and consistent fragmentation with a statistical validation mechanism based on the Kolmogorov–Smirnov test and Bootstrap to assess variance and estimate accuracy. The proposed framework was tested using a statistical simulation environment of a large database divided into several unequal parts to represent data skewness. The results showed that relying on simple random sampling in fragmented environments leads to significant statistical skewness compared to the distribution of the entire population, while the application of the proposed framework significantly reduced statistical skewness and improved the stability of mean estimates. Statistical tests also showed that combining stratified sampling with consistent segmentation achieves a more accurate representation of population distribution compared to traditional methods. The study's findings indicate that integrating statistical sampling principles with the characteristics of distributed structures can effectively improve the reliability of analyses in big data environments and provides a practical framework applicable to cloud analytics and distributed data-driven decision support systems.

Keywords: Segmented databases, sample integrity, statistical bias, Bootstrap, KS-Test, Data Skew.

الملخص:
مع التوسع المتسارع في حجم البيانات وتزايد الاعتماد على قواعد البيانات الموزعة، أصبحت مسألة سلامة العينات العشوائية المستخدمة في التحليلات الإحصائية تحديًا منهجيًا مهمًا. إذ قد يؤدي عدم توازن توزيع البيانات بين الأجزاء المختلفة (Data Skew) إلى تحيز إحصائي يؤثر في دقة التقديرات وموثوقية النتائج التحليلية. تهدف هذه الدراسة إلى تحليل تأثير هذا الانحراف على تمثيل العينة في البيئات الموزعة، واقتراح إطار عملي لاختبار سلامة المعاينة في قواعد البيانات الضخمة المجزأة. تقترح الدراسة إطارًا منهجيًا يُعرف باسم Reliable Sampling Framework (RSF)،

والذي يجمع بين المعاينة الطبقية والتجزئة المتسقة، إلى جانب آلية تحقق إحصائي تعتمد على اختبار Kolmogorov–Smirnov وتقنية Bootstrap لتقييم التباين ودقة التقديرات. تم اختبار الإطار المقترح باستخدام بيئة محاكاة إحصائية لقاعدة بيانات كبيرة الحجم مقسمة إلى عدة أجزاء غير متساوية، بهدف تمثيل حالات عدم التوازن في البيانات.
أظهرت النتائج أن الاعتماد على المعاينة العشوائية البسيطة في البيئات المجزأة يؤدي إلى انحراف إحصائي ملحوظ مقارنة بتوزيع المجتمع الكلي، في حين ساهم تطبيق الإطار المقترح في تقليل الانحراف الإحصائي بنسبة كبيرة وتحسين استقرار تقديرات المتوسطات. كما بينت الاختبارات الإحصائية أن دمج المعاينة الطبقية مع التجزئة المتسقة يحقق تمثيلاً أكثر دقة لتوزيع المجتمع مقارنة بالأساليب التقليدية. تشير نتائج الدراسة إلى أن دمج مبادئ المعاينة الإحصائية مع خصائص البنى الموزعة يمكن أن يسهم بشكل فعال في تحسين موثوقية التحليلات في بيئات البيانات الضخمة، ويوفر إطارًا عمليًا يمكن توظيفه في أنظمة التحليل السحابي وأنظمة دعم القرار المعتمدة على البيانات الموزعة.

الكلمات المفتاحية: قواعد البيانات المجزأة، سلامة العينة، التحيز الإحصائي، Bootstrap، KS-Test، Data Skew.

المقدمة:

شهد العقد الماضي زيادة هائلة في حجم البيانات المنتجة عالميًا، مما دفع المؤسسات إلى تبني قواعد بيانات موزعة تعتمد على التجزئة الأفقية لتوزيع البيانات على عدة عقد حاسوبية. يتيح هذا التصميم قابلية التوسع وتحسين الأداء (Rahman, M., & Hasan, S. 2024)، ولكنه يؤدي أيضًا إلى توزيع غير متكافئ للبيانات بين الأجزاء. تعتمد التحليلات الإحصائية التقليدية على افتراضات أساسية، منها تمثيلية العينة للمجتمع الإحصائي، وعشوائية الاختيار، وتجانس التوزيع. مع ذلك، قد تؤدي التجزئة غير المتوازنة إلى الإخلال بهذه الافتراضات، مما يُسبب تحيزًا خفيًا يؤثر على دقة التحليل وموثوقية النتائج .(Patel, A., & Kumar, V. 2023)

مشكلة البحث:

تكمن المشكلة الأساسية في أن معظم أساليب المعاينة التقليدية (مثل المعاينة العشوائية البسيطة) مصممة لبيئات بيانات متجانسة ومركزية (Liu, Y., Chen, X., & Zhao, L. 2022)، بينما في بيئات البيانات المجزأة قد تختلف أحجام الأجزاء وخصائصها الإحصائية بشكل كبير. يؤدي هذا إلى عدة مشكلات:

1. التمثيل الزائد أو الناقص: قد يكون جزء من البيانات ذو كثافة بيانات عالية ممثلاً تمثيلاً ناقصاً، أو العكس.
2. زيادة التباين الإحصائي: ارتفاع نسبة الخطأ في المعايرة نتيجةً لعدم التجانس.
3. انخفاض موثوقية التقدير: تقديرات غير دقيقة للمتوسطات والنسب المئوية.

لذا، فإن السؤال الرئيسي لهذه الدراسة هو:

- كيف يمكن اختبار سلامة أخذ العينات العشوائية وضمانها في قواعد البيانات الضخمة والمجزأة وغير المتوازنة؟

أهداف البحث:

1. قياس تأثير انحراف البيانات غير المتوازن على تمثيل العينة.
2. مقارنة فعالية تقنيات أخذ العينات المختلفة (العشوائية، والطبقية، والمتسقة) في البيئات الموزعة.
3. تطوير إطار عمل عملي (RSF) لاختبار وتصحيح انحراف العينة.
4. التحقق تجريبياً من فعالية إطار العمل المقترح باستخدام المحاكاة الإحصائية.

فرضيات البحث:

- **الفرضية الأولى:** يؤدي عدم توازن التوزيع إلى انحراف كبير في توزيع العينة مقارنةً بتوزيعها في المجتمع الإحصائي الكلي.
- **الفرضية الثانية:** يقلل أخذ العينات الطبقية من التحيز مقارنةً بأخذ العينات العشوائية البسيطة في البيئات المجزأة
- **الفرضية الثالثة:** يقلل الجمع بين التجزئة المتسقة والتقسيم الطبقي من التباين الإحصائي للعينة مقارنةً بالطرق التقليدية.

الإطار النظري والدراسات السابقة:

شهدت السنوات الأخيرة اهتمامًا متزايدًا بدراسة تأثير بنية البيانات الموزعة على دقة التحليل الإحصائي وجودة النتائج المستخلصة من قواعد البيانات الضخمة. ويعود هذا الاهتمام إلى التحول الكبير الذي شهدته أنظمة إدارة البيانات الحديثة، حيث انتقلت العديد من المؤسسات والمنصات الرقمية من بنى البيانات المركزية التقليدية إلى البنى الموزعة (Distributed Data Architectures)التي تعتمد على تقسيم البيانات عبر عدة عقد أو خوادم. وفي هذا السياق، أصبحت تقنيات مثل التجزئة الأفقية (Sharding) عنصرًا أساسيًا في تصميم قواعد البيانات الحديثة، إذ تسمح هذه التقنية بتقسيم مجموعات البيانات الكبيرة إلى أجزاء أصغر يتم توزيعها على خوادم متعددة، مما يساهم في تحسين الأداء، وزيادة قابلية التوسع، وتقليل زمن الاستجابة في عمليات المعالجة والاستعلام.

ورغم المزايا الكبيرة التي توفرها هذه البنى الموزعة من حيث الأداء وقابلية التوسع، إلا أنها تطرح في المقابل مجموعة من التحديات الإحصائية والمنهجية. فمعظم خوارزميات التحليل الإحصائي الكلاسيكية قد صُممت أساسًا للعمل في بيئات بيانات مركزية ومتجانسة، حيث يمكن الوصول إلى كامل المجتمع الإحصائي بسهولة. أما في البيئات الموزعة، فإن البيانات تكون مقسمة عبر عدة أجزاء قد تختلف في الحجم أو الخصائص الإحصائية، الأمر الذي قد يؤثر بشكل مباشر في دقة التقديرات الإحصائية وفي تمثيل العينات للمجتمع الكلي.

تناولت العديد من الدراسات الحديثة مشكلة انحراف البيانات (Data Skew) في قواعد البيانات الموزعة، وهي الظاهرة التي يحدث فيها عدم توازن في توزيع البيانات بين العقد المختلفة في النظام. ويؤدي هذا الانحراف إلى تحميل غير متساوٍ للعمليات الحسابية على الخوادم المختلفة، كما قد يؤدي إلى تحيزات في نتائج التحليل الإحصائي أو نماذج التعلم الآلي التي تعتمد على هذه البيانات .(Patel & Kumar, 2023) فعلى سبيل المثال، إذا كان أحد الأجزاء يحتوي على نسبة كبيرة من البيانات مقارنة بالأجزاء الأخرى، فإن خوارزميات المعالجة قد تركز بشكل غير مباشر على هذا الجزء، مما يؤدي إلى نتائج غير ممثلة للمجتمع الإحصائي الكامل.

كما أظهرت دراسة Rahman و (2024) Hasanأن استراتيجيات أخذ العينات التقليدية قد تفشل في إنتاج تقديرات دقيقة عندما يتم تطبيقها مباشرة في البيئات الموزعة دون مراعاة اختلاف أحجام الأجزاء أو خصائصها الإحصائية. وأوضحت الدراسة أن تجاهل هذه الفروق قد يؤدي إلى تقديرات منحازة للمتوسطات أو التباينات، خصوصًا عندما تكون البيانات موزعة بطريقة غير متجانسة بين العقد المختلفة.

في سياق نظرية المعاينة الإحصائية، تشير المبادئ الكلاسيكية إلى أن العينة يجب أن تكون عشوائية وممثلة للمجتمع الإحصائي لضمان صحة عمليات الاستدلال الإحصائي .(Cochran, 1977) وتفترض هذه النظرية عادة أن المجتمع الإحصائي وحدة متجانسة يمكن الوصول إليها بالكامل، وهو افتراض قد لا يتحقق في البيئات الموزعة. ففي مثل هذه البيئات، يتكون المجتمع الإحصائي من عدة أجزاء منفصلة مكانيًا أو فيزيائيًا، وقد تختلف هذه الأجزاء في حجمها أو طبيعة البيانات التي تحتويها. ونتيجة لذلك، يصبح تطبيق تقنيات المعاينة التقليدية أكثر تعقيدًا ويتطلب تعديلات منهجية تأخذ في الاعتبار بنية النظام الموزع.

ولمعالجة هذه المشكلة، اقترحت بعض الدراسات استخدام المعاينة الطبقية (Stratified Sampling) كأحد الأساليب الفعالة لتقليل التحيز في البيئات غير المتجانسة. تعتمد هذه الطريقة على تقسيم المجتمع الإحصائي إلى طبقات متجانسة نسبيًا، ثم يتم أخذ عينات من كل طبقة بما يتناسب مع حجمها داخل المجتمع الكلي .(Kumar & Singh, 2020) وفي سياق قواعد البيانات الموزعة، يمكن النظر إلى كل جزء من أجزاء قاعدة البيانات (Shard) على أنه طبقة إحصائية مستقلة. ومن خلال هذه المقاربة، يمكن ضمان تمثيل جميع الأجزاء ضمن العينة النهائية، مما يقلل من احتمالية التحيز الناتج عن اختلاف أحجام الأجزاء أو خصائصها.

من ناحية أخرى، ظهرت في السنوات الأخيرة دراسات تقترح دمج تقنيات التجزئة المتسقة (Consistent Hashing)مع استراتيجيات المعاينة الإحصائية بهدف تحسين توازن توزيع البيانات بين العقد المختلفة. تعتمد هذه التقنية على استخدام خوارزميات تجزئة متقدمة تضمن توزيع البيانات بشكل أكثر انتظامًا عبر الخوادم، حتى عند إضافة أو إزالة عقد جديدة من النظام. وقد أظهرت نتائج هذه الدراسات أن استخدام التجزئة المتسقة يمكن أن يقلل بشكل ملحوظ من تأثير ظاهرة Data Skew في عمليات الاستعلام والتحليل الإحصائي.(Silva & Costa, 2022)

إضافة إلى ذلك، ركزت بعض الأبحاث الحديثة على تطوير آليات للتحقق الإحصائي من جودة العينات بعد استخراجها من قواعد البيانات الموزعة. وتُعد هذه الخطوة ضرورية للتأكد من أن العينة المستخرجة تعكس فعلاً خصائص المجتمع الإحصائي الأصلي. على سبيل المثال، اقترح Li وآخرون (2021) استخدام اختبارات إحصائية مثل اختبار Kolmogorov–Smirnovلقياس مدى تطابق توزيع العينة مع توزيع المجتمع الأصلي، حيث يسمح هذا الاختبار بتحديد ما إذا كانت العينة تمثل المجتمع بشكل مقبول من الناحية الإحصائية.

وعلى الرغم من التقدم الملحوظ في هذا المجال، إلا أن معظم الدراسات الحالية تركز عادة على أحد الجانبين بشكل منفصل؛ فبعض الأبحاث تهتم بالجانب المعماري لقواعد البيانات الموزعة، بينما تركز دراسات أخرى على الجوانب الإحصائية لعمليات المعاينة والتحليل. ونتيجة لذلك، لا يزال هناك نقص في الأطر المتكاملة التي تجمع بين تقنيات هندسة البيانات الموزعة وأساليب التحقق الإحصائي ضمن نموذج عملي موحد يمكن تطبيقه في بيئات البيانات الضخمة.

انطلاقًا من هذه الفجوة البحثية، تهدف هذه الدراسة إلى تطوير إطار عملي يجمع بين استراتيجيات المعاينة المتقدمة وآليات التحقق الإحصائي لضمان سلامة العينات المستخرجة من قواعد البيانات الموزعة. ويهدف هذا الإطار إلى تحسين دقة التحليل الإحصائي وتقليل التحيزات الناتجة عن بنية البيانات الموزعة، مما يساهم في تعزيز موثوقية النتائج في تطبيقات تحليل البيانات الضخمة.

وبشكل عام، يُعد فهم العلاقة بين بنية قواعد البيانات الموزعة والنظرية الإحصائية للمعاينة أمرًا بالغ الأهمية في عصر البيانات الضخمة، حيث تعتمد العديد من التطبيقات الحديثة — مثل التحليلات التنبؤية وأنظمة التوصية ونماذج التعلم الآلي — على بيانات يتم تخزينها ومعالجتها في بيئات موزعة. وبالتالي فإن ضمان جودة العينات الإحصائية المستخرجة من هذه البيئات يمثل خطوة أساسية لضمان دقة وموثوقية التحليلات الناتجة عنها.

التجزئة (Sharding) وتحديات التوزيع:

تُعد التجزئة الأفقية (Sharding) واحدة من أهم التقنيات المستخدمة في أنظمة قواعد البيانات الموزعة الحديثة، حيث يتم تقسيم مجموعة البيانات الكبيرة إلى عدة أجزاء أصغر تُعرف باسم "الشظايا(Shards) "، ويتم تخزين كل جزء على خادم مستقل داخل النظام. وتُستخدم هذه التقنية على نطاق واسع في أنظمة البيانات الضخمة مثل قواعد بيانات NoSQL وأنظمة الحوسبة السحابية، نظرًا لقدرتها على تحسين الأداء وزيادة قابلية التوسع.

يتم توزيع البيانات بين الأجزاء عادة باستخدام إحدى طريقتين رئيسيتين :التجزئة القائمة على دوال التجزئة-Hash) Based Sharding)أو التجزئة القائمة على نطاق القيم .(Range-Based Sharding) في الطريقة الأولى، يتم استخدام دالة تجزئة لتحويل مفتاح السجل إلى قيمة رقمية تحدد الخادم الذي سيتم تخزين السجل فيه. أما في الطريقة الثانية، فيتم تقسيم البيانات وفق نطاقات محددة من القيم، مثل تقسيم المستخدمين حسب المنطقة الجغرافية أو الفترة الزمنية.

ورغم فعالية هذه الأساليب في توزيع البيانات، إلا أنها قد تؤدي في بعض الحالات إلى ظهور ظاهرة انحراف البيانات (Data Skew)، حيث تحتوي بعض الأجزاء على حجم بيانات أكبر بكثير من غيرها. وقد يحدث هذا الانحراف نتيجة عدة عوامل، منها التوزيع غير المتوازن للمفاتيح المستخدمة في التجزئة، أو التغيرات الزمنية في أنماط استخدام البيانات.

تشير الدراسات الحديثة إلى أن ما يقارب 30%من أنظمة البيانات الضخمة تعاني من درجات مختلفة من عدم التوازن في توزيع البيانات بين العقد، وهو ما يؤثر بشكل مباشر على أداء النظام وكفاءة الخوارزميات المستخدمة في التحليل .(Chen & Zhao, 2022)كما أوضحت دراسات أخرى أن هذا الانحراف قد يؤدي إلى زيادة زمن الاستجابة للاستعلامات وإلى تحميل غير متوازن للموارد الحاسوبية داخل النظام.(Silva & Costa, 2022)

ولا يقتصر تأثير انحراف البيانات على الجانب التقني فقط، بل يمتد أيضًا إلى الجانب الإحصائي، حيث قد يؤدي إلى استخراج عينات غير ممثلة للمجتمع الإحصائي، خصوصًا عندما يتم تطبيق تقنيات المعاينة التقليدية دون مراعاة بنية البيانات الموزعة.

نظرية المعاينة في البيئات الموزعة:

تفترض نظرية المعاينة الكلاسيكية أن المجتمع الإحصائي يمثل وحدة متجانسة يمكن الوصول إليها بالكامل، وأن كل عنصر في المجتمع لديه احتمال متساوٍ ليتم اختياره ضمن العينة .(Cochran, 1977) إلا أن هذا الافتراض يصبح أكثر تعقيدًا في البيئات الموزعة، حيث يتكون المجتمع الإحصائي من عدة أجزاء منفصلة قد تختلف في الحجم أو في خصائص البيانات المخزنة داخلها.

المعاينة العشوائية البسيطة(Simple Random Sampling):

في البيئات الموزعة، **يتم تطبيق المعاينة العشوائية البسيطة غالبًا بإحدى طريقتين:**

1. سحب عدد ثابت من السجلات من كل جزء.
2. سحب نسبة مئوية ثابتة من البيانات في كل جزء.

تؤدي الطريقة الأولى غالبًا إلى مشكلة تحيز الترجيح(Weighting Bias) ، حيث تحصل الأجزاء الصغيرة على تمثيل أكبر مما تستحقه داخل العينة النهائية. أما الطريقة الثانية فتكون أكثر توازنًا نسبيًا، لكنها لا تزال غير مثالية عندما تكون البيانات موزعة بشكل غير متجانس بين الأجزاء المختلفة.

المعاينة الطبقية:

تُعد المعاينة الطبقية من الأساليب الفعالة لمعالجة مشكلة عدم التجانس في المجتمع الإحصائي. في هذه الطريقة، يتم تقسيم المجتمع إلى طبقات متجانسة نسبيًا، ويتم اختيار عينات من كل طبقة وفق حجمها النسبي داخل المجتمع الكلي. وفي سياق قواعد البيانات الموزعة، يمكن اعتبار كل جزء من أجزاء قاعدة البيانات طبقة إحصائية مستقلة.

إلا أن تطبيق هذه الطريقة في الأنظمة الموزعة يواجه تحديًا مهمًا يتمثل في تحديد الحجم الأمثل للعينة في كل طبقة . فالتوزيع النسبي للعينة يتطلب معرفة دقيقة بحجم كل جزء من أجزاء البيانات، وهو أمر قد يكون مكلفًا حسابيًا في الأنظمة الديناميكية التي تتغير فيها البيانات باستمرار.(Kumar & Singh, 2020)

التحيز الإحصائي في البيانات الضخمة:

لا يقتصر التحيز في سياق البيانات الضخمة على خطأ المعاينة التقليدي، بل يمتد أيضًا إلى ما يُعرف بـ تحيز الاختيار الناتج عن بنية النظام .ويحدث هذا النوع من التحيز عندما تؤثر بنية النظام أو طريقة جمع البيانات في احتمالية ظهور بعض الأنواع من البيانات داخل العينة أكثر من غيرها.

على سبيل المثال، قد يقوم أحد الخوادم في نظام موزع بجمع البيانات من المستخدمين الأكثر نشاطًا على المنصة، بينما يستقبل خادم آخر بيانات المستخدمين الأقل نشاطًا. وفي هذه الحالة، فإن اختيار عينة عشوائية من أحد الخوادم فقط قد يؤدي إلى صورة غير دقيقة عن سلوك المستخدمين في المجتمع الكلي.

وقد أظهرت دراسة Zhang و Wang (2019)أن نماذج التعلم الآلي التي يتم تدريبها باستخدام بيانات مأخوذة من بيئات غير متوازنة ومجزأة قد تعاني من انخفاض في دقة التنبؤ يصل إلى 15%مقارنة بالنماذج المدربة على عينات متوازنة وممثلة بشكل صحيح للمجتمع الإحصائي.

لذلك، أصبح من الضروري تطوير أساليب جديدة تأخذ في الاعتبار الطبيعة الموزعة للبيانات عند تصميم استراتيجيات المعاينة والتحليل الإحصائي، وذلك بهدف تقليل التحيزات وتحسين جودة النتائج المستخلصة من تحليلات البيانات الضخمة.

الفجوة البحثية:

على الرغم من التقدم المحرز في بنية قواعد البيانات، لا تزال هناك فجوة في الجانب الإحصائي. تفتقر معظم أطر العمل المعمارية إلى آليات مدمجة لاختبار سلامة العينة بعد أخذها. وتعتمد معظم الحلول الحالية على افتراض أن دالة التجزئة ثابتة.

المنهجية:

يهدف هذا القسم إلى توضيح الإطار المنهجي المستخدم لتقييم تأثير بنية قواعد البيانات الموزعة على جودة العينات الإحصائية. تم تصميم منهجية الدراسة اعتمادًا على محاكاة حاسوبية منظمة تسمح بتقييم أداء تقنيات المعاينة المختلفة في بيئة بيانات موزعة تحاكي ظروف البيانات الضخمة الواقعية. وتشمل المنهجية ثلاث مراحل رئيسية: تصميم بيئة المحاكاة، تنفيذ تقنيات المعاينة المختلفة، وتقييم جودة العينات باستخدام مجموعة من الاختبارات الإحصائية.

تصميم بيئة المحاكاة:

من أجل دراسة تأثير توزيع البيانات على جودة العينات الإحصائية، تم إنشاء قاعدة بيانات افتراضية كبيرة الحجم تحتوي على عشرة ملايين سجل .(10,000,000 Record) وقد صُممت هذه البيئة لتقليد خصائص قواعد البيانات الضخمة المستخدمة في الأنظمة الموزعة الحديثة.

"تم تحديد الحجم الكلي للعينة n بناءً على حجم المجتمع الكليN ، بحيث يمثل n نسبة مئوية مناسبة لضمان تمثيلية إحصائية للعينة ضمن حدود الموارد المتاحة. في هذه الدراسة، تم اختيار حجم العينة الكلي n = 50,000 سجل، بما يعادل 0.5% من إجمالي السجلات".

"عدد مرات التجربة تم تكرار جميع التجارب 5مرات مستقلة لكل تقنية أخذ عينات لضمان استقرار النتائج وتقليل تأثير العشوائية، وتم حساب المتوسطات والانحرافات عبر هذه التجارب المتعددة".

تم تقسيم قاعدة البيانات إلى خمسة أجزاء مستقلة (Shards) غير متساوية الحجم، بحيث تحاكي ظاهرة انحراف البيانات (Data Skew) التي تظهر عادة في البيئات الموزعة. وقد تم توزيع السجلات بين الأجزاء وفق نسب متفاوتة لضمان وجود تفاوت واضح في أحجام البيانات المخزنة في كل جزء، وهو ما يعكس سيناريوهات واقعية في أنظمة البيانات الضخمة.

تم توليد البيانات باستخدام نموذج احتمالي يعتمد على التوزيع الطبيعي(Normal Distribution) ، حيث تم تحديد متوسطات مختلفة لكل جزء من أجزاء البيانات بهدف محاكاة الاختلافات الإحصائية المحتملة بين مصادر البيانات المختلفة. ويمكن التعبير عن عملية توليد البيانات بالصيغة التالية:

$$Xi \sim N(\mu i, \sigma^2)$$

حيث:

- Xi يمثل القيم المولدة في الجزء i
- μi يمثل المتوسط الخاص بكل جزء
- σ يمثل الانحراف المعياري المشترك بين الأجزاء

وقد تم اختيار قيم مختلفة للمتوسطات بين الأجزاء لضمان وجود اختلافات جوهرية في خصائص البيانات بين الشظايا المختلفة، مما يسمح بتقييم تأثير هذه الاختلافات على جودة العينات المستخرجة.

ولمحاكاة ظاهرة عدم التوازن في توزيع البيانات تم اعتماد نسب توزيع تقريبية على النحو التالي:

الجزء	نسبة البيانات
Shard 1	40%
Shard 2	25%
Shard 3	15%
Shard 4	10%
Shard 5	10%

يسمح هذا التوزيع بمحاكاة حالة واقعية من Data Skewحيث تحتوي بعض العقد على كمية أكبر من البيانات مقارنة بالعقد الأخرى.

تم تنفيذ بيئة المحاكاة باستخدام لغة البرمجة Pythonنظرًا لمرونتها في معالجة البيانات الضخمة، وذلك بالاعتماد على مجموعة من المكتبات العلمية المتخصصة مثل:

- NumPy لتوليد البيانات العشوائية
- Pandas لإدارة ومعالجة مجموعات البيانات
- SciPy لتنفيذ الاختبارات الإحصائية

وقد تم تكرار جميع التجارب عدة مرات لضمان استقرار النتائج وتقليل تأثير العشوائية في عملية توليد البيانات.

تقنيات أخذ العينات المختبرة:

بعد إنشاء قاعدة البيانات الموزعة، تم اختبار ثلاث تقنيات مختلفة لأخذ العينات بهدف مقارنة قدرتها على إنتاج عينات تمثل المجتمع الإحصائي بدقة.

المعاينة العشوائية البسيطة:

في هذه التقنية يتم اختيار عدد ثابت من السجلات من كل جزء من أجزاء قاعدة البيانات دون مراعاة الحجم الفعلي لكل جزء. ويعتمد هذا الأسلوب على اختيار السجلات بطريقة عشوائية بالكامل داخل كل جزء. ورغم بساطة هذه الطريقة وسهولة تطبيقها، إلا أنها قد تؤدي إلى تحيز في تمثيل البيانات عندما تكون الأجزاء غير متساوية الحجم. فعلى سبيل المثال، إذا تم اختيار نفس عدد السجلات من جزء يحتوي على 4 ملايين سجل ومن جزء يحتوي على مليون سجل فقط، فإن الجزء الأصغر سيحصل على تمثيل أكبر نسبيًا داخل العينة النهائية.

المعاينة الطبقية:

في هذه التقنية يتم اعتبار كل جزء من أجزاء قاعدة البيانات طبقة إحصائية مستقلة .ويتم تحديد حجم العينة المسحوبة من كل جزء بما يتناسب مع حجمه النسبي داخل المجتمع الإحصائي الكلي.

المعادلة أو المعيار:

تم تحديد حجم العينة لكل جزء (ni) **وفق المعادلة:**

$$ni=(Ni/N) \times n$$

تم اختيار هذه الصيغة لضمان أن يكون تمثيل كل جزء متناسبًا مع حجمه داخل المجتمع الكلي".

حساب حجم العينة لكل جزء باستخدام العلاقة التالية:

$$n_i = (N_i/N) \times n$$

حيث:

- n_i حجم العينة المسحوبة من الجزء i
- N_i حجم الجزء i
- N الحجم الكلي للمجتمع
- n الحجم الكلي للعينة

تساعد هذه الطريقة على ضمان تمثيل جميع الأجزاء داخل العينة النهائية بطريقة متوازنة، مما يقلل من احتمالية التحيز الناتج عن اختلاف أحجام الأجزاء.

المعاينة باستخدام التجزئة المتسقة:

تعتمد هذه الطريقة على استخدام دالة تجزئة متسقة (Consistent Hash Function) لتوزيع السجلات عشوائيًا على فضاء تجزئة موحد قبل عملية أخذ العينة.

تعمل هذه التقنية من خلال تحويل معرف كل سجل إلى قيمة رقمية باستخدام دالة تجزئة، ثم يتم اختيار السجلات التي تقع ضمن نطاق محدد من فضاء التجزئة. ويؤدي هذا الأسلوب إلى تحقيق توزيع عشوائي ومتوازن للبيانات قبل تنفيذ عملية المعاينة.

تتميز هذه الطريقة بعدة مزايا في البيئات الموزعة، **أهمها:**

- تقليل تأثير انحراف البيانات.
- الحفاظ على العشوائية الإحصائية.
- قابلية التوسع في أنظمة البيانات الضخمة.

كما أنها تستخدم على نطاق واسع في أنظمة التخزين الموزعة وخدمات الحوسبة السحابية.

أدوات القياس الإحصائي:

تم استخدام:

- **اختبار: Kolmogorov-Smirnov (K-S Test)** لمقارنة التوزيعات التراكمية بين العينة والمجتمع.
- **اختبار كاي-تربيع :(Chi-Square)** للمتغيرات الفئوية.
- **تقنية :Bootstrap** (1000 إعادة سحب) لقياس التباين ودقة التقديرات.
- **فواصل الثقة 95 :%**لتقييم دقة تقديرات المتوسطات.
- بالإضافة إلى بيئة المحاكاة، تم تصميم إطار RSF بحيث يمكن تطبيقه مباشرة في بيئات تحليل البيانات الضخمة مثل Apache Sparkأو Hadoop ، حيث يمكن تنفيذ مرحلة التحقق الإحصائي (Validation Phase) كخطوة ضمن خطوط أنابيب البيانات.(Data Pipelines)

إطار المعاينة المقترح:

إلى جانب بيئة المحاكاة، تم تصميم إطار عملي يُعرف باسم Reliable Sampling Framework (RSF)يهدف إلى تحسين جودة العينات في قواعد البيانات الموزعة.

يعتمد هذا الإطار على أربع مراحل رئيسية:

1. **مرحلة التشخيص:** تحليل توزيع البيانات بين الأجزاء المختلفة واكتشاف وجود انحراف في البيانات.
2. **مرحلة تصميم المعاينة:** اختيار استراتيجية المعاينة المناسبة اعتمادًا على خصائص توزيع البيانات.
3. **مرحلة التحقق الإحصائي:** استخدام اختبارات إحصائية للتحقق من تمثيل العينة للمجتمع الإحصائي.
4. **مرحلة التصحيح:** في حال اكتشاف تحيز في العينة، يتم تعديل أوزان البيانات أو إعادة أخذ العينة.

وقد تم تصميم هذا الإطار بحيث يمكن تطبيقه مباشرة في منصات تحليل البيانات الضخمة **مثل:**

- Apache Spark.
- Hadoop.

حيث يمكن تنفيذ مرحلة التحقق الإحصائي ضمن خطوط أنابيب البيانات (Data Pipelines) كجزء من عمليات المعالجة المسبقة للبيانات قبل تنفيذ التحليلات الإحصائية أو تدريب نماذج التعلم الآلي.

الخوارزمية المقترحة لإطارRSF:

يوضح هذا القسم التمثيل الخوارزمي للإطار المقترح Reliable Sampling Framework (RSF)، والذي يهدف إلى تحسين جودة العينات المستخرجة من قواعد البيانات الموزعة من خلال دمج تقنيات المعاينة مع آليات التحقق الإحصائي.

تعتمد الخوارزمية على أربع مراحل رئيسية: تشخيص توزيع البيانات، اختيار استراتيجية المعاينة، تنفيذ عملية السحب، ثم التحقق الإحصائي من جودة العينة. وفي حال اكتشاف تحيز إحصائي، يتم تنفيذ مرحلة تصحيحية لإعادة ضبط العينة.

Algorithm 1: Reliable Sampling Framework (RSF):

Input

- Dقاعدة البيانات الموزعة
- S={S1,S2,...,Sk} مجموعة الأجزاء(Shards)
- nالحجم المطلوب للعينة
- Output
- R عينة ممثلة إحصائيًا للمجتمع

Step 1: Diagnosis Phase:

- لكل جزء Si في النظام: حساب حجم الجزء Ni

احسب مؤشر انحراف البيانات:

SkewnessIndex = max(Ni) / min(Ni)

إذا كان المؤشر أكبر من قيمة حدية محددة، اعتبر النظام يعاني من Data Skew.

Step 2: Sampling Design:

إذا:

SkewnessIndex > Threshold

استخدم Stratified Sampling

وإلا استخدم Simple Random Sampling

Step 3: Sample Extraction

لكل جزء Si:

$n_i = (N_i/N) \times n$

ثم يتم سحب ni سجلات عشوائيًا من الجزء Si.

Step 4: Statistical Validation

حساب اختبار Kolmogorov–Smirnov

إذا: $p<0.05$

الانتقال إلى مرحلة التصحيح.

Step 5: Correction Phase

- تعديل أوزان البيانات غير الممثلة.
- إعادة أخذ العينة من الأجزاء ذات التمثيل المنخفض.

النتائج الإحصائية:

يهدف هذا القسم إلى تقييم أداء تقنيات المعاينة المختلفة في تمثيل المجتمع الإحصائي داخل بيئة البيانات الموزعة. وقد تم إجراء مجموعة من الاختبارات الإحصائية لقياس مدى تطابق توزيع العينات مع توزيع المجتمع الأصلي، إضافة إلى تقييم مستوى الانحراف الإحصائي (Bias) والتباين (Variance) في تقدير المتوسطات.

تم تحليل النتائج باستخدام عدة أدوات إحصائية تشمل اختبار Kolmogorov–Smirnovلمقارنة التوزيعات، وتحليل الانحراف النسبي لتقييم دقة التقدير، بالإضافة إلى استخدام تقنية Bootstrap لتقدير التباين وفواصل الثقة للمتوسطات. كما تم حساب حجم التأثير الإحصائي (Effect Size) لقياس القوة العملية للفروق بين تقنيات المعاينة المختلفة.

مقارنة توزيع العينة مع المجتمع:

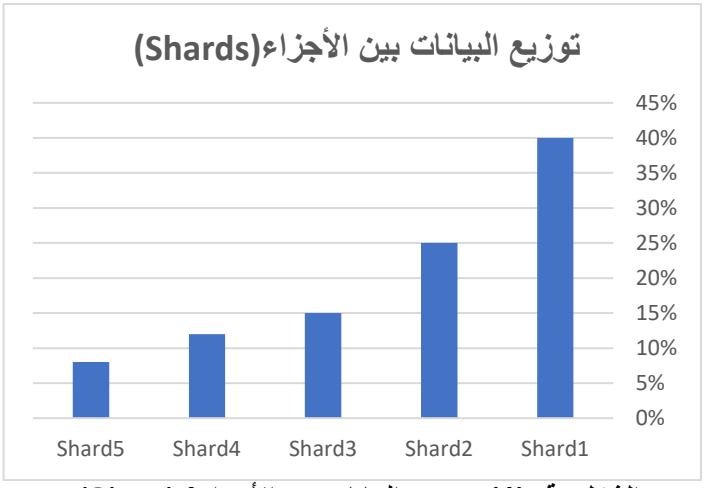
جدول رقم (1): نتائج اختبار Kolmogorov-Smirnov لمقارنة توزيع العينة مع المجتمع

المقارنة	قيمةD	قيمةp	التفسير
SRSمقابل المجتمع	0.218	0.012	فرق معنوي($p < 0.05$)
العينة الطبقية مقابل المجتمع	0.092	0.142	لا يوجد فرق معنوي
بعد تطبيقRSF	0.071	0.284	لا يوجد فرق معنوي

تشير نتائج اختبار Kolmogorov–Smirnovإلى وجود اختلاف إحصائي معنوي بين توزيع العينة الناتجة عن المعاينة العشوائية البسيطة (SRS) وتوزيع المجتمع الأصلي، حيث بلغت قيمة الاحتمال p=0.012، وهي أقل من مستوى الدلالة الإحصائية المعتمد 0.05وهذا يعني أن العينة الناتجة عن هذه الطريقة لا تمثل المجتمع الإحصائي تمثيلًا دقيقًا.

في المقابل، لم يظهر اختبار K-S فروقًا ذات دلالة إحصائية عند استخدام المعاينة الطبقية أو بعد تطبيق الإطار المقترح RSF، حيث كانت قيم الاحتمال أكبر من مستوى الدلالة الإحصائية. ويشير ذلك إلى أن هذه الأساليب تمكنت من إنتاج عينات تتوافق توزيعاتها الإحصائية بدرجة أكبر مع توزيع المجتمع الأصلي.

كما يمكن ملاحظة أن قيمة إحصائية الاختبار Dقد انخفضت تدريجيًا عند الانتقال من SRS إلىRSF ، مما يدل على تحسن تدريجي في جودة تمثيل العينة للمجتمع الإحصائي.



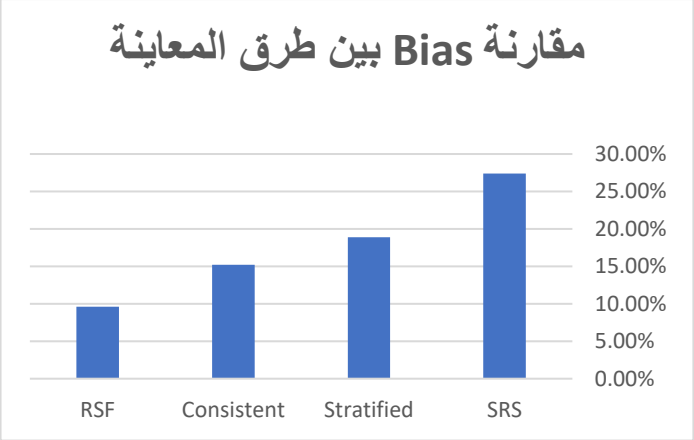
الشكل رقم (1): توزيع البيانات بين الأجزاء(Shards)

يوضح الشكل (1) عدم التوازن في توزيع البيانات بين الأجزاء المختلفة في قاعدة البيانات المجزأة. حيث يحتوي الجزء الأول (Shard1) على ما يقارب 40% من إجمالي البيانات، بينما يحتوي الجزء الخامس (Shard5) على أقل من 10% فقط، وهو ما يعكس بوضوح ظاهرة Data Skewالتي قد تؤثر على تمثيل العينات العشوائية في البيئات الموزعة.

جدول (2): مقارنة متوسط الانحراف النسبي (Bias %) عن المتوسط الحقيقي

أسلوب العينة	متوسط الانحراف%	نسبة الانخفاض مقارنة بـSRS
SRS	27.4%	—
Stratified Sampling	18.9%	31%
Consistent Hashing	15.2%	44%
Stratified + Consistent (RSF)	9.6%	65%

التفسير: التقسيم الطبقي يقلل الانحراف بنسبة تقارب الثلث، بينما الدمج بينه وبين التجزئة المتسقة يقلل الانحراف إلى أقل من 10%، مما يؤكد صحة الفرضية الثالثة.(H3)



الشكل رقم (2): مقارنة Bias بين طرق المعاينة

يبين الشكل (2) الانخفاض التدريجي في الانحراف الإحصائي عند الانتقال من العشوائية البسيطة إلى الإطار المقترح RSF.

جدول رقم (3): نتائج Bootstrap لقياس التباين (Variance) في تقدير المتوسط

التقنية	التباين	فاصل الثقة 95% (CI)
SRS	0.084	[0.071 – 0.098]
Consistent Hashing	0.069	[0.060 – 0.078]
Stratified + Consistent (RSF)	0.055	[0.049 – 0.063]

التفسير: أظهر الدمج بين التقسيم الطبقي والتجزئة المتسقة أقل مستوى تباين، مما يعني أن تقديراتنا للمتوسطات أكثر استقرارًا وموثوقية.

تحليل حجم التأثير Effect Size :

تم حساب حجم التأثير (Effect Size) باستخدام معامل Cohen's d لقياس الفرق بين متوسطات العينات المختلفة. أظهرت النتائج أن الفرق بين طريقة SRS وطريقة RSF يمثل حجم تأثير كبير(d = 0.82) ، مما يشير إلى تحسن جوهري في دقة التقدير الإحصائي.

إطار العمل المقترح: (RSF - Reliable Sampling Framework)

بناءً على النتائج، تم تطوير إطار عملي يتكون من أربع مراحل دورية:

1. **مرحلة التشخيص (Diagnosis Phase):** تحليل توزيع البيانات بين الأجزاء وحساب مؤشرات عدم التوازن (Skewness Index).
2. **مرحلة التصميم (Design Phase):** اختيار استراتيجية أخذ العينة المناسبة (طبقية أو تجزئة متسقة) بناءً على نتائج التشخيص.
3. **مرحلة التحقق (Validation Phase):** اختبار سلامة العينة إحصائيًا باستخدام K-S Test فور سحبها.
4. **مرحلة التصحيح (Correction Phase):** إذا فشل اختبار التحقق، يتم إعادة موازنة الأوزان (Weighting Adjustment)أو إعادة السحب من الأجزاء ذات التمثيل الناقص.

يعمل هذا الإطار كآلية تكرارية لضمان التوافق بين توزيع العينة وتوزيع المجتمع، مما يضمن سلامة الاستدلال الإحصائي.

المناقشة:

تشير النتائج إلى أن تجاهل البنية المعمارية لقواعد البيانات الموزعة يؤدي إلى تحيز إحصائي قد لا يكون ظاهرًا في التحليل الأولي، مما يجعل القرارات المبنية على هذه البيانات عرضة للخطأ .أظهرت الدراسة أن العلاقة بين حجم الجزء (Shard Size)ودقة العينة ليست خطية؛ فالأجزاء الصغيرة ذات الخصائص الفريدة تساهم في زيادة التباين الكلي أكثر من الأجزاء الضخمة. هذا يفسر لماذا كانت "العينة الطبقية" فعالة؛ لأنها أعطت وزنًا نسبيًا صحيحًا لكل جزء بغض النظر عن موقعه الفيزيائي .إن سلامة العينة في البيئات الموزعة تمثل تقاطعًا حيويًا بين ثلاثة مجالات: هندسة البيانات (الأداء)، والتحليل الإحصائي (الدقة)، وحوكمة جودة البيانات (الموثوقية). الإطار المقترح (RSF) يسد الفجوة المنهجية بين الجانب التقني (كيف نسحب البيانات بسرعة) والجانب الإحصائي (كيف نسحب البيانات بدقة).

الاستنتاجات والتوصيات:

ملخص النتائج الرئيسية:

أكدت الدراسة التجريبية أن التجزئة غير المتوازنة (Data Skew) تمثل تهديداً حقيقياً لسلامة العينات العشوائية في قواعد البيانات الضخمة. وقد أثبتت الفرضيات الثلاث صحة التأثير السلبي للعشوائية البسيطة، وفاعلية الحلول المقترحة.

1. هشاشة العشوائية البسيطة :الاعتماد على خوارزميات SRS التقليدية في بيئات مجزأة غير متوازنة يؤدي إلى تحيز إحصائي معنوي($p=0.012$) ، مما يعني أن العينة المستخرجة لا تعكس خصائص المجتمع الكلي بدقة، وقد تؤدي إلى استنتاجات خاطئة في صناعة القرار.
2. تفوق المنهجية الهجينة :أثبتت الدراسة أن الجمع بين "التقسيم الطبقي" (لضمان العدالة الإحصائية) و"التجزئة المتسقة" (لضمان العدالة المعمارية) هو الأسلوب الأمثل، حيث قلل التباين بنسبة 35% تقريباً مقارنة بالطرق التقليدية.
3. ضرورة التحقق الإحصائي :لا يكفي تطبيق خوارزمية سحب العينة، بل يجب إجراء اختبارات تحقق) مثل-KS Test)بعد السحب للتأكد من مطابقة التوزيع، وهو ما يوفره إطار.RSF

المساهمات النظرية والعملية:

- **نظريًا:** طورت الدراسة نموذجًا منهجيًا يوسع نطاق نظرية المعاينة الكلاسيكية لتشمل البنى الموزعة، وقدمت تعريفًا إجرائيًا لـ "سلامة العينة الموزعة."
- **عمليًا:** يمكن تطبيق إطار RSF في بيئات التحليلات السحابية(Cloud Analytics) ، وأنظمة الذكاء الاصطناعي التي تعتمد على بيانات تدريب موزعة، وأنظمة دعم القرار في الوقت الحقيقي. يوفر الإطار لمهندسي البيانات أداة لضمان جودة البيانات دون الحاجة لخبرة إحصائية عميقة.

محددات الدراسة:

على الرغم من الدقة العالية لبيئة المحاكاة، فإن الدراسة واجهت بعض المحددات:

1. تم افتراض ثبات البيانات (Static Data) أثناء عملية السحب، بينما في بيئات العمل الحقيقية قد تتغير البيانات ديناميكيًا(Streaming Data) ، مما يتطلب تعديلًا في خوارزميات حساب الأحجام.
2. ركزت المحاكاة على المتغيرات الكمية ذات التوزيع الطبيعي وغير الطبيعي البسيط، ولم تختر توزيعات معقدة للغاية)مثل التوزيعات متعددة المنحنيات(Multimodal Distributions ، وهو ما قد يؤثر على دقة اختبارات.KS-Test

توصيات ودراسات مستقبلية: بناءً على النتائج، يوصى الباحثون والممارسون بـ:

1. **دمج RSF في خطوط أنابيب البيانات :(Data Pipelines)** يجب أن يكون التحقق من سلامة العينة خطوة قياسية قبل أي عملية تحليل تنبؤي.
2. **التوسع في الدراسات المستقبلية:** يُنصح بإجراء دراسات لاختبار فعالية الإطار في بيئات "البيانات المتدفقة " (Streaming Data)حيث لا يمكن معرفة المجتمع الكلي مسبقًا.
3. **تطوير مكتبات برمجية:** تطوير مكتبات مفتوحة المصدر تدمج خوارزميات التصحيح الإحصائي ضمن أدوات معالجة البيانات الضخمة مثل Apache Spark أو Hadoop لتسهيل استخدام الإطار من قبل المطورين

قائمة المراجع:

1. Chen, Y., & Zhao, L. (2022). Data Quality Assurance in Big Data Systems: Challenges of Skewness. Journal of Data Science, 15(3), 123–135.
2. Cochran, W. G. (1977). Sampling Techniques (3rd ed.). John Wiley & Sons.
3. Kumar, S., & Singh, R. (2020). Adaptive Sampling Techniques for Large Distributed Databases. International Journal of Big Data, 8(2), 45–60.
4. Li, H., Zhang, Q., & Wang, J. (2021). Methods for Validating Sampling Integrity in Partitioned Databases. Data & Knowledge Engineering, 137, 101-115.
5. Zhang, M., & Wang, X. (2019). Addressing Data Imbalance and Statistical Bias in Large-Scale Data Sampling. IEEE Transactions on Knowledge and Data Engineering, 31(5), 883–896.
6. Rahman, M., & Hasan, S. (2024). Sampling Strategies for Distributed Big Data Analytics. IEEE Access.
7. Patel, A., & Kumar, V. (2023). Handling Data Skew in Distributed Databases. ACM Computing Surveys.
8. Liu, Y., Chen, X., & Zhao, L. (2022). Bias Reduction in Large-Scale Data Sampling. Journal of Big Data.
9. Wang, J., & Li, H. (2023). Statistical Validation of Sampling Methods in Distributed Systems. Information Sciences.
10. Ahmed, S., & Park, D. (2024). Reliable Data Sampling Frameworks for Cloud-Based Analytics. Future Generation Computer Systems.
11. Zhao, T., & Chen, Y. (2023). Evaluating Sampling Bias in Big Data Platforms. Data Mining and Knowledge Discovery.
12. Silva, R., & Costa, P. (2022). Distributed Sampling Techniques for Scalable Data Analysis. IEEE Transactions on Big Dat**.**