الإطار النظري للبيانات الضخمة وتقنياتها

بنظرة سريعة على مراحل التطور التكنولوجي لنظم معالجة البيانات الرقمية يتبين لنا أن ظهور البيانات الضخمة مر بثلاثة مراحل أساسية تتمثل فيما يلي:

- ✓ المرحلة الأولى: الانتقال من الثقافة الورقية إلى الأوعية الرقمية؛
- ✓ المرحلة الثانية: تم الانتقال من أجهزة الكمبيوتر إلى شبكة الأنترنت؛
- ✓ المرحلة الثالثة: تم الانتقال فيها من شبكة الأنترنت إلى البيانات الضخمة.

2-مفهوم البيانات الضخمة وأهميتها:

تعرف مؤسسة البيانات الدولية (IDC) تقنية البيانات الضخمة على أنها "جيل جديد من التقنيات المصممة لاستخراج القيمة الاقتصادية من كميات كبيرة جدًا ومتنوعة من البيانات، من خلال الالتقاط السريع والاكتشاف والتحليل".
يسهل الفهم الشامل للمؤسسة ويطور الاستراتيجيات وخرائط الطريق للمؤسسات، خاصةً عندما يتعلق الأمر بمشاركة العملاء والأتمتة والتحليلات التنبؤية التي تساعد في اتخاذ القرارات؛

- ✓ يساعد على تحسين ذكاء الأعمال وتحديد أهم البيانات التي تفيد قطاع الأعمال وتحسين كفاءة ودقة التحليلات والتوقعات وتوجيه القرارات المستقبلية؛
- ✓ تساعد تحليلات البيانات الضخمة الشركات على فهم سلوك العملاء بشكل أفضل من أجل الوصول لرضاهم، وتساعد كذلك في تحديد كيفية زيادة المعلومات الاستخباراتية وتحسين كفاءة الأعمال، وزيادة المبيعات وفرص الابتكار والتنبؤ بنمو المبيعات والأرباح...الخ.

3-خصائص البيانات الضخمة:

يمكن استعراض أهم الخصائص التي تميز البينات الضخمة فيما يلي:

3-1-الحجمvolume : تتميز البينات الضخمة بكبر حجمها وهو ما يفسر احتياجها لأجهزة وبرامج متطورة لها القدرة على التعامل مع هذا الكم الهائل من البيانات، كما أن حجم البيانات هو الذي يحدد قيمتها وإمكانياتها لكي تصنف على أنها بيانات ضخمة، والذي يقاس بالبيتا بايت والاكسا بايت.

3-2-التنوع variety : تتنوع البيانات الضخمة ما بين بيانات مهيكلة والتي تكون منظمة في جداول أو قواعد بيانات، وبيانات غير مهيكلة مثل مقاطع الفيديو والصور وتسجيلات الصوت وبيانات الخرائط، وغيرها من البيانات التي تتطلب وقتا وجهدا لتهيئتها في شكل مناسب يسمح بتحليلها.

3-3-السرعة Velocity: يقصد بها الوقت الذي يتم استغراقه في إنتاج واستخراج البيانات لتغطية الطلب عليها، حيث تكون الأنظمة قادرة على إنشاء البيانات بسرعة عالية جدا، وذلك من أجل تخزينها ومعالجتها وتحليلها بشكل آني.

3-4-الموثوقية والحجة Veracity: يقصد بها مدى دقة البيانات وحداثتها وصحتها، أي تحري مدى موثوقية مصدرها.

3-5-القيمة Value : تمثل قيمة البيانات الفوائد والرؤى المستقبلية التي يمكن أن تولدها للشركات، حيث تمثل هذه الخاصية الهدف الأساسي لجمع وتحليل كم ضخم وهائل من البيانات،

4-أنواع البيانات الضخمة:

تصنف البيانات الضخمة إلى ثلاث أنوع من البيانات هي:

4-1-البيانات المهيكلة: وهي بيانات تكون منظمة في شكل جداول (في الغالب)، تتميز بأنها سهلة للبحث فيها وكذلك سهلة المعالجة، حيث يمكن معالجتها باستعمال أدوات معالجة البيانات التقليدية.

4-2-البيانات غير الهيكلية: وهي بيانات غير منظمة تمثل النسبة الأكبر من البيانات الضخمة، تكون في شكل صور ومقاطع فيديو ورسوم بيانية وصور، أي تتمثل في كل مالا يمكن تنظيمه أو تصنيفه بسهولة، ويتطلب هذا النوع من البيانات أدوات وتقنيات متطورة مناسبة لمعالجتها وتحليلها.

4-3-البيانات شبه المهيكلة: تصنف هذه البيانات على أنها بيانات مهيكلة إلا أنها لا تكون في شكل قواعد بيانات ولا تكون منظمة في شكل جداول مثل برامج معالجة النصوص.

5-مصادر البيانات الضخمة: تناولت العديد من الدراسات المصادر الأساسية التي تقوم عليها البيانات الضخمة والتي تعتبر من الأسباب الأساسية لكبر حجمها وضخامتها كما قلنا، ويمكن تصنيفها كما يلي:

5-1-مصادر داخلية للبيانات: وتتعلق هذه المصادر بالنظم والعمليات الداخلية للشركة التي تمتلك الأجهزة المادية والإلكترونية وتساهم في معالجة وحفظ ومشاركة البيانات بين تلك الأجهزة عبر الأنترنت، وتتمثل تلك النظم في نظم إدارة علاقات العملاء Customer Relations Management (CRM)، ونظم تخطيط موارد المنشأة Enterprise Resource planning (ERP)، ونظم أنترنت الأشياء(IOT)، ونظم الذكاء الاصطناعي Artificial (AI) Intelligence وغيرها من النظم.

5-2-مصادر خارجية للبيانات: وتتعلق بالبيانات التي لا تمتلكها الشركات إلا أنها تتأثر بها وتؤثر فيها ومن بينها: بيانات السوق والمنافسين، بالإضافة إلى البيانات المتأتية من شبكات الأنترنت ومواقع التواصل الاجتماعي التي أصبحت عنصر هاما يؤثر في بيئة الأعمال ومن أمثلتها: Facebook, Anstigram, Twitter، فضلا عن بيانات الأقمار الصناعية GPS، وسجلات تداول الأوراق المالية والبيانات التي تحتويها بطاقات الائتمان.

- **برامج ومنصات تحليل البيانات الضخمة:**

إن برامج وأساليب التنقيب عن البيانات التقليدية غير ملائمة تمامًا للطبيعة المعقدة للبيانات الضخمة (الحجم، السرعة والتنوع)، فغالبًا ما تكون هذه البرامج والأساليب غير فعّالة وبطيئة وتفشل في ضمان تخزين كميات كبيرة غير متجانسة من البيانات ومعالجتها تحليلها بفعالية مع السرعة المطلوبة، وهذه المتطلبات تم التوصل إليها بفضل تطور تكنولوجيا الحوسبة وظهور التقنيات الجديدة، وفيما يلي نظرة عامة على التقنيات الجديدة التي قدمتها ثورة البيانات الضخمة:

1- منصة Apache Hadoop : Hadoop هو برنامج يديره Apache ، ويعتبر من أشهر أدوات وأساليب تحليل البيانات الضخمة، وهو مفتوح المصدر يعمل بلغة الجافا ويقوم بمعالجة وتخزين البيانات الضخمة بأشكال مختلفة مثل تخزين بيانات ضخمة على أجهزة متعددة ومعالجتها ومن ثم توزيع عملية المعالجة على

هذه الأجهزة لتسريع النتيجة المتوصل اليها من عملية المعالجة، ويعمل Hadoop على تنفيذ نموذج حسابي يسمى Reduce Map يقوم بالإرسال إلى جميع الخوادم/ السيرفرات في نفس التوقيت، وكل سيرفر أو خادم يعطي ما لديه من بيانات، بعدها يتم تجميع هذه البيانات وإرسالها كحزمة واحدة، ومن أشهر مستخدمي الهادوب Hadoop : Amazon, Apple, AVG, eBay, Electronic Arts, Facebook, Google, IBM, LinkedIn, Microsoft, the New York Times, Twitter, Yahoo.

2-**منصة Microsoft HDInsight**: طُرحت هذه المنصة من طرف شركة ميكروسوفت، تعمل هذه المنصة على إدارة ومعالجة وتحليل البيانات الضخمة بجودة عالية وتكلفة منخفضة، وتستخدم هذه المنصة نظام Azure Windows Blob كنظام الملفات الافتراضية.

3--منصة HPCC: هي منصة سهلة الاستخدام ومجانية لأنها مفتوحة المصدر صممت لمعالجة البيانات الضخمة بجودة وأداء عالي.

4-**منصة PolyBase**: صممت هذه المنصة لمعالجة البيانات الضخمة العلائقية*، حيث أنها تعمل من خلال مستودع بيانات(SQL Server 2012 Parallel Data Warehouse (PDW، ولها القدرة على التكامل مع منصة Hadoop ما يسمح لها بالوصول إلى جميع البيانات بما فيها غير العلائقية أيضا.

5-**برنامج EXCEL**: هي أداة من أدوات Microsoft شائعة الاستعمال ويمكن استخدامها في تحليل ومعالجة البيانات الضخمة من خلال ربط البيانات المخزنة في Hadoop.

6-**منصة OpenRefine**: وتعرف كذلك باسم منصة Google Refine، وُضعت هذه المنصة خصيصا للتعامل مع البيانات غير المهيكلة أو غير المنظمة، حيث تعمل على تجميع البيانات من مواقع ويب متنوعة وتحليلها وتحويل صيغتها.

4-التحديات التي تواجه البيانات الضخمة:

أجمعت كل الدراسات على مجموعة من التحديات التي يمكن أن تواجه المنظمات عند استخدمها للبيانات الضخمة يمكن عرضها على النحو التالي:

- ✓ **التحديات المرتبطة بالبيانات:** كالنمو الهائل والمتسارع في حجم البيانات الضخمة، وتمّيز هذه البيانات بضخامة الحجم، والتنوع، والسرعة والخصوصية؛
- ✓ **التحديات المرتبطة بالمعالجة:** كعدم توفر الموظفين والكوادر المتخصصة في استخدام التقنيات الحديثة لتحليل هذه البيانات، بالإضافة إلى صعوبة الاختيار بين البيانات المتشابهة، وتحويلها إلى شكل قابل للتحليل، ونمذجة البيانات وفهم المخرجات، وكيفية عرض التحليلات المعقدة للبيانات؛
- ✓ **التحديات المرتبطة بالإدارة**: وتشمل خصوصية هذه البيانات، أمنها، حوكمتها وأخلاقياتها.